Video: Protein-Protein Interactions
Credit: Mauro Martino

Network Science

Cristian Candia-Castro Vallejos, Ph.D.

cristiancandia@udd.cl

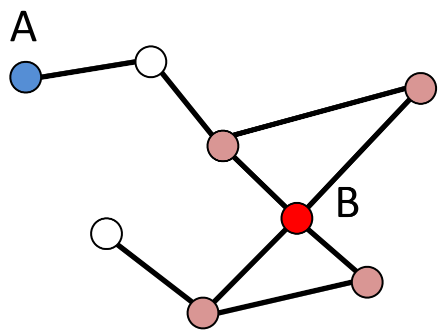
Director Magister en Data Science UDD
Profesor Investigador, Facultad de Ingeniería, UDD
External Faculty Northwestern Institute on Complex Systems,
Kellogg School of Management, Northwestern University

Modelos de Formación de Redes

Estas diapositivas se basan en la presentación original del
Prof. Albert-László Barabási, de Northeastern University, con autorización.
El contenido ha sido traducido para su uso en este curso.

Distribución de grado

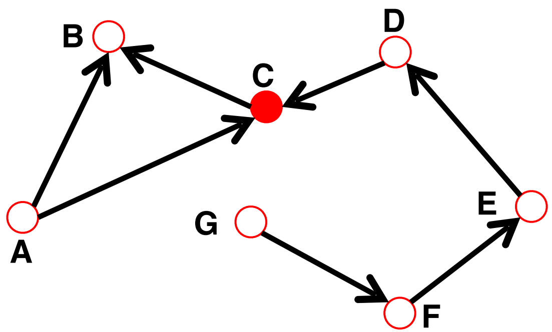
No-dirigido



Grado del nodo: el número de links conectados al nodo.

$$k_A = 1 \qquad k_B = 4$$

Dirigido



En ***redes dirigidas*** Podemos definer un grado de entrada (in-degree) y un grado de salida (out-degree). El grado total es la suma de ambos.

$$k_C^{in} = 2 \qquad k_C^{out} = 1 \qquad k_C = 3$$

Fuente: un nodo con $k^{in}= 0$; Sumidero: un nodo con $k^{out}= 0$.

Estadística

Breve revisión estadística

Four key quantities characterize a sample of N values $x_1, \dots, x_N$:

Promedio:

$$\langle x \rangle = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

El n-esimo momento:

$$\langle x^n \rangle = \frac{x_1^n + x_2^n + \ldots + x_N^n}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i^n$$

Desviación estándar

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \langle x \rangle\right)^2}$$

Distribución de *x*

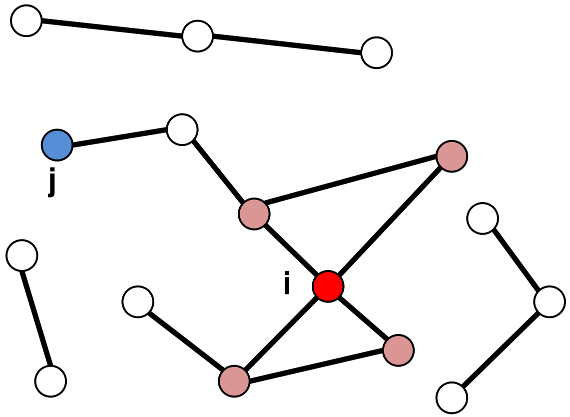
$$p_x = \frac{1}{N}\sum_i \delta_{x,x_i}$$

Donde *x* sigue:

$$\sum_i p_x = 1 \left(\int p_x \, dx = 1\right)$$

Grado promedio

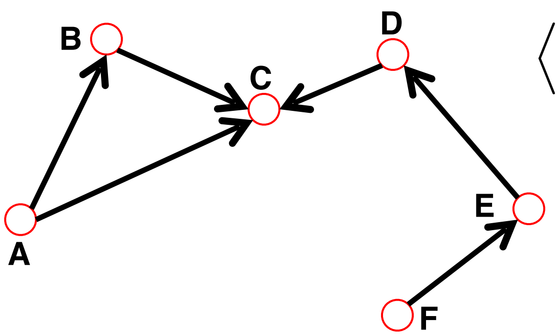
No-dirigido



$$\langle k \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i \qquad \langle k \rangle \equiv \frac{2L}{N}$$

N – el número de nodos en el grafo

Dirigido



$$\langle k^{in} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{in}, \quad \langle k^{out} \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i^{out}, \quad \langle k^{in} \rangle = \langle k^{out} \rangle$$

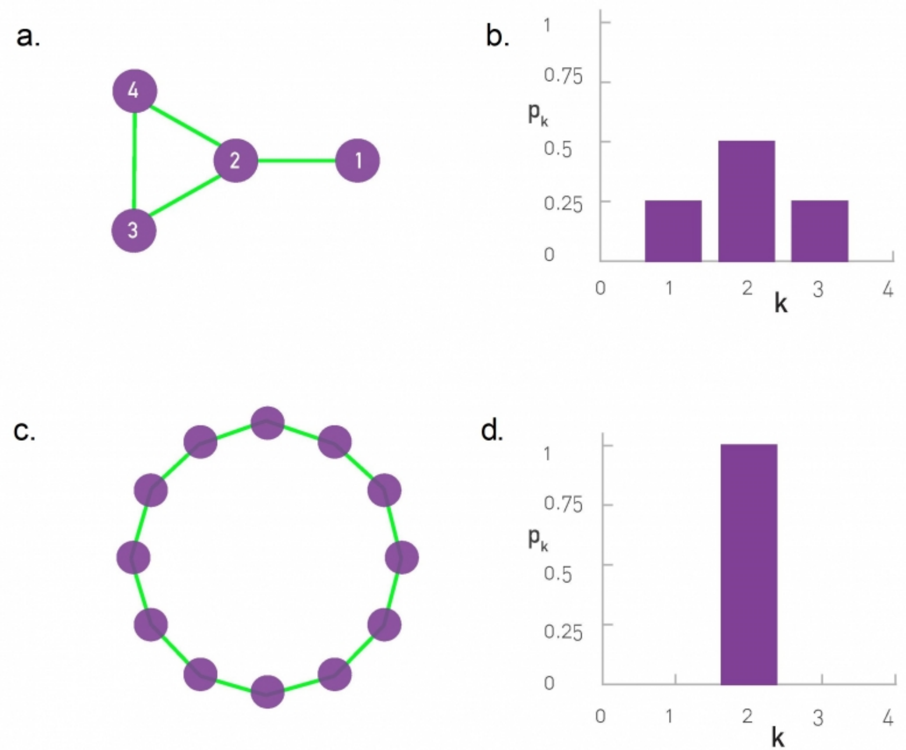
$$\langle k \rangle \equiv \frac{L}{N}$$

Grado promedio

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L	⟨k⟩
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.33
WWW	Webpages	Links	Directed	325,729	1,497,134	4.60
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594	2.67
Mobile Phone Calls	Subscribers	Calls	Directed	36,595	91,826	2.51
Email	Email addresses	Emails	Directed	57,194	103,731	1.81
Science Collaboration	Scientists	Co-authorship	Undirected	23,133	93,439	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Paper	Citations	Directed	449,673	4,689,479	10.43
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802	5.58
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930	2.90

Distribución de grado

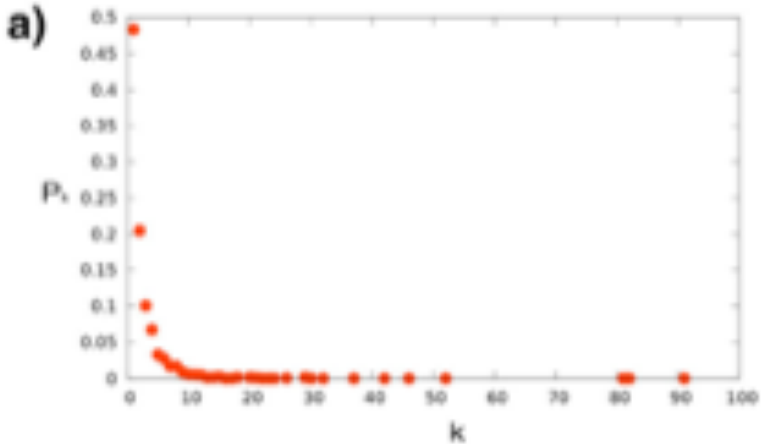
P(k): probabilidad de que un nodo elegído aleatoriamente tenga grado *k*



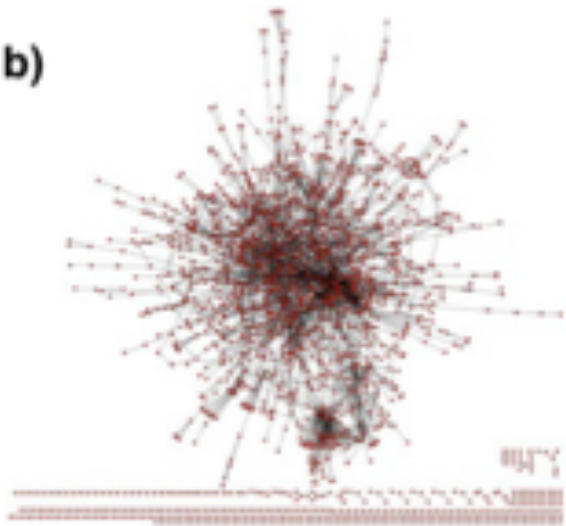
N_k = # nodos con grado k

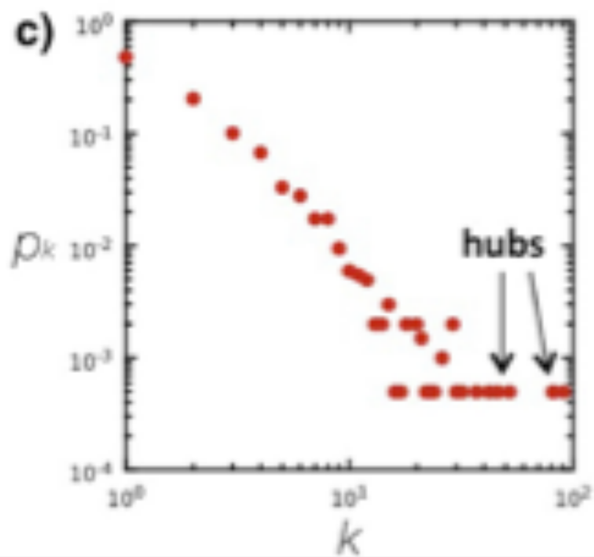
$P(k) = N_k / N$ → plot

Distribución de grado



En muchas redes reales, el grado de nodo puede variar considerablemente.
Por ejemplo, como indica la distribución de grados (a), los grados de las proteínas en la red de interacción de proteínas que se muestran en (b) varían entre k = 0 (nodos aislados) y k = 92, que es el grado del nodo más grande, llamado un centro. También hay grandes diferencias en el número de nodos con diferentes grados: como muestra (a), casi la mitad de los nodos tienen grado uno (es decir, p1 = 0,48), mientras que solo hay una copia del nodo más grande, por lo tanto, p_92 = 1 / N = 0.0005. (c) La distribución de grados a menudo se muestra en el llamado gráfico log-log, en el que trazamos log p_k en función de log k, o, como hicimos en (c), usamos ejes logarítmicos.





DISTRIBUCION DE GRADO

Representación discreta: $\mathbf{p_k}$ Es la probabilidad de que un nodo tenga grado **k**.

Descripción continua: **p(k)** es la pdf de los grados, donde

$$\int_{k_1}^{k_2} p(k)\,dk$$

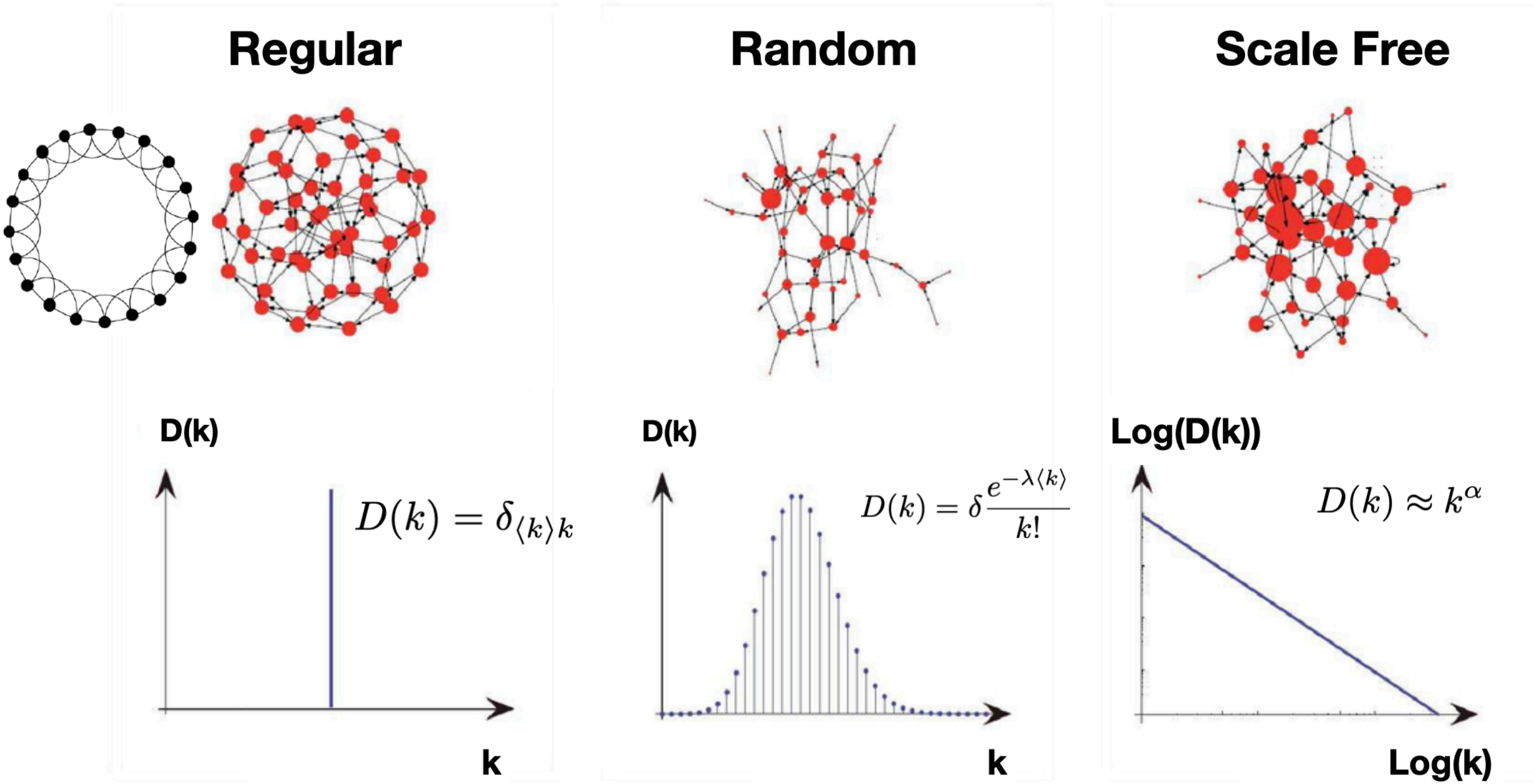
representa la probabilidad de que el grado de un nodo esté entre $\mathbf{k_1}$ y $\mathbf{k_2}$.

Condición de normalización:

$$\sum_{0}^{\infty} p_k = 1 \qquad\qquad \int_{K_{\min}}^{\infty} p(k)\,dk = 1$$

donde K_{min} es el mínimo grado de la red.

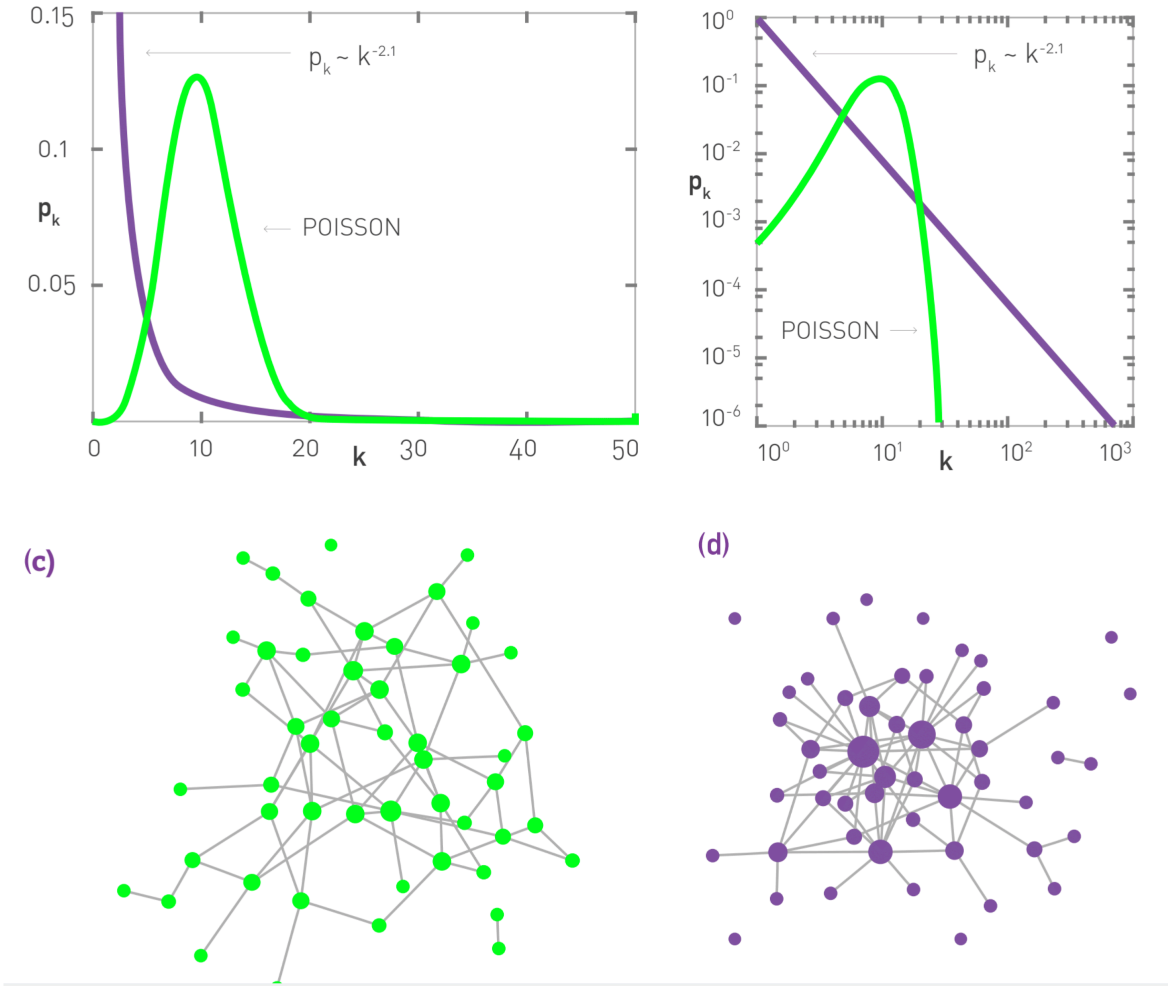
Distribución de grado



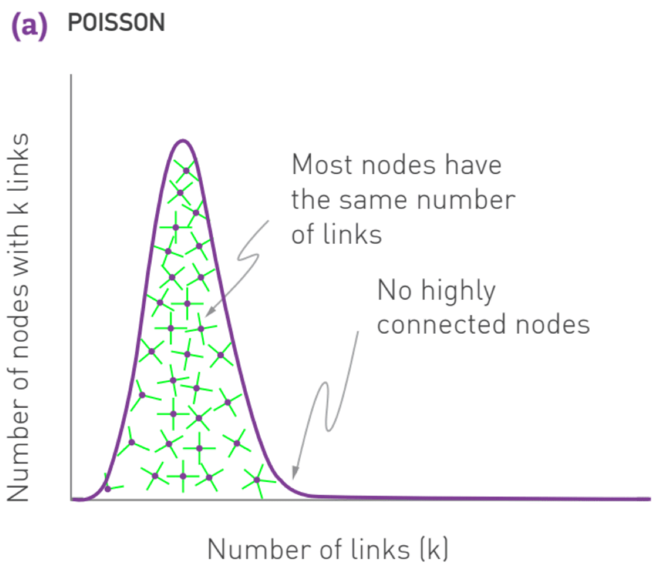
Piensen en los histogramas de fercuencias de grado

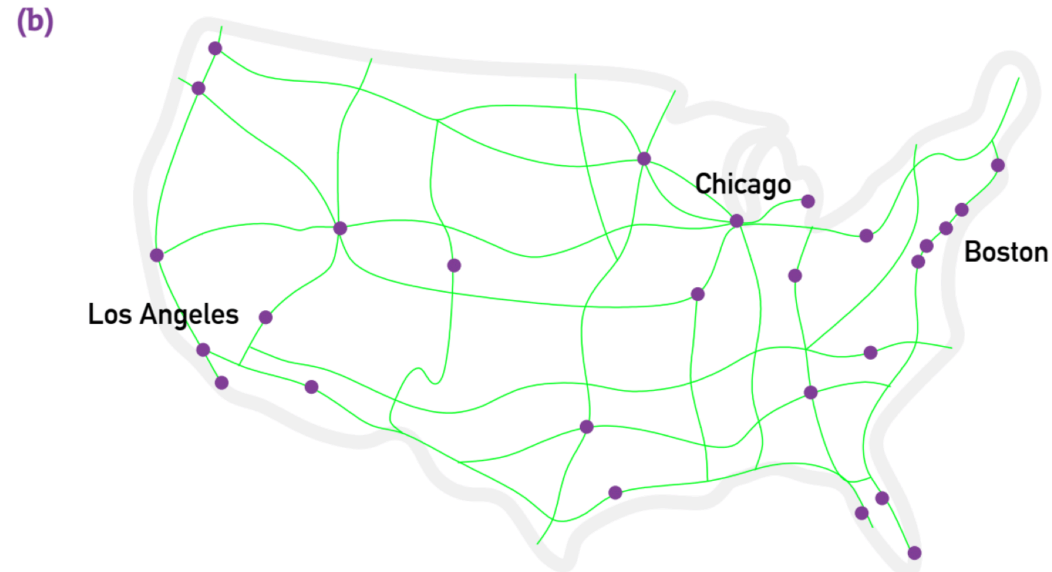
¿Cómo se comparan los promedios con los valores máximos?

Diferencia entre red aletoria y libre de escala

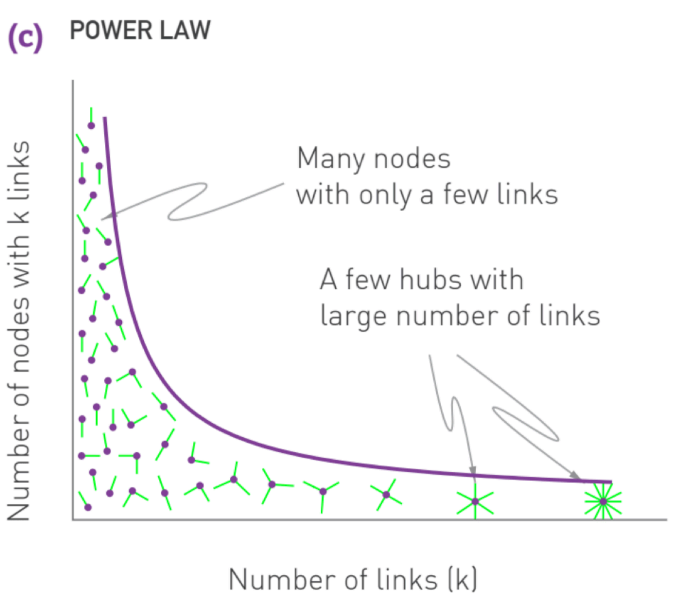


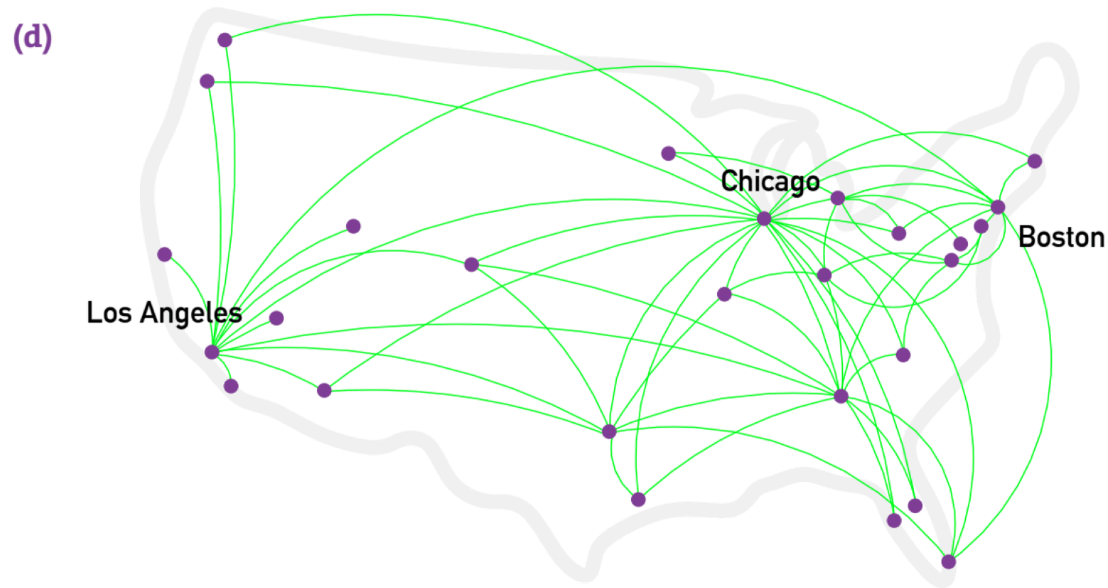
Diferencia entre red aletoria y libre de escala





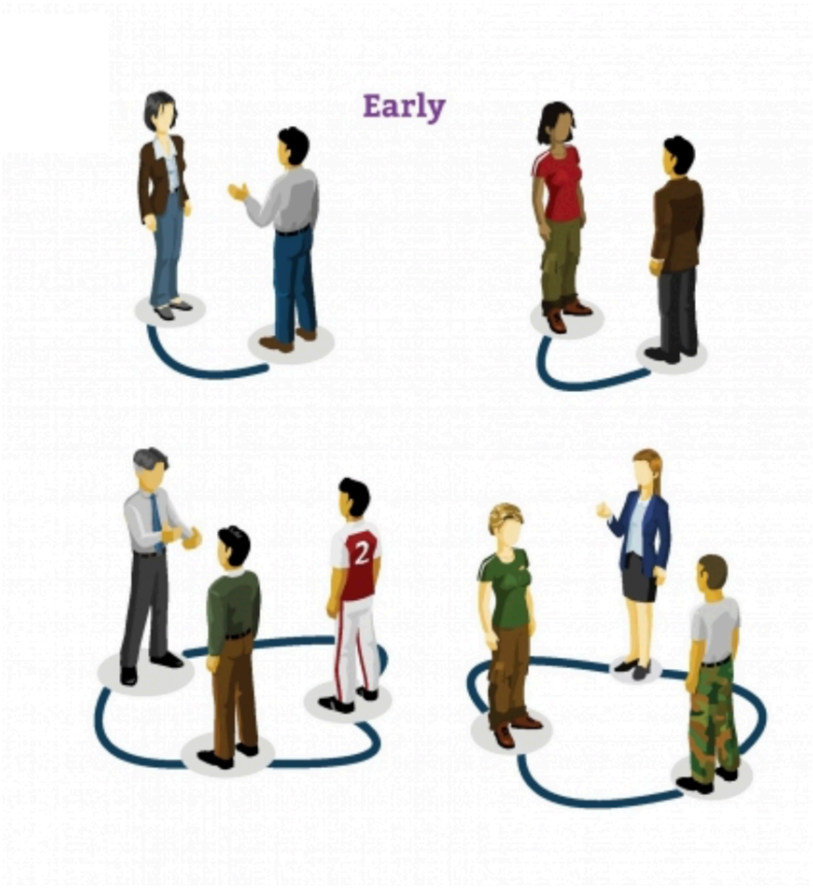
Diferencia entre red aletoria y libre de escala



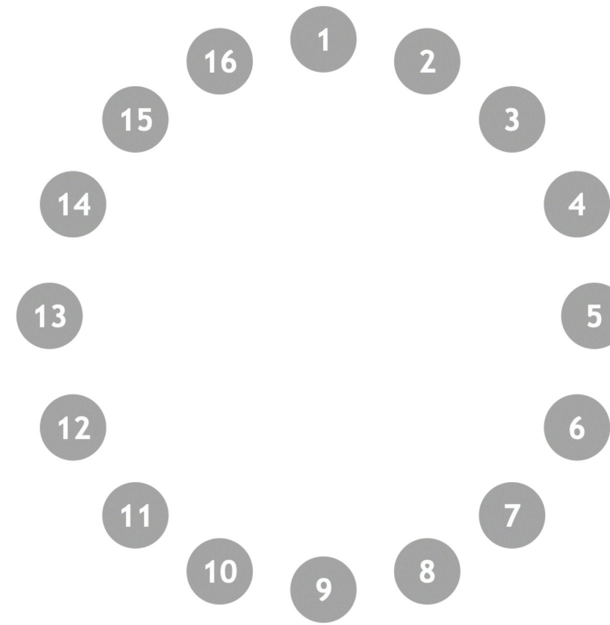


Estas estructuras tienen un impacto significativo en cómo se propaga la información en los sistemas físicos, biológicos, sociales, etc.

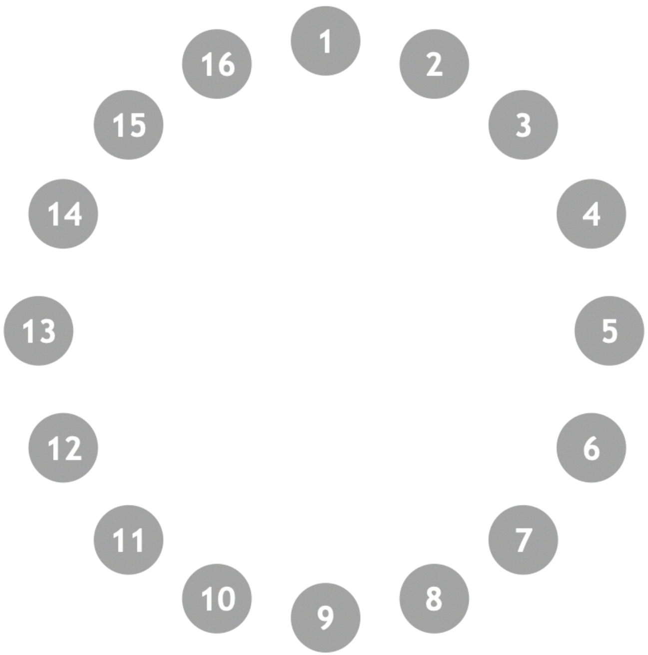
MODELO DE REDES ALEATORIAS



the party algorithm

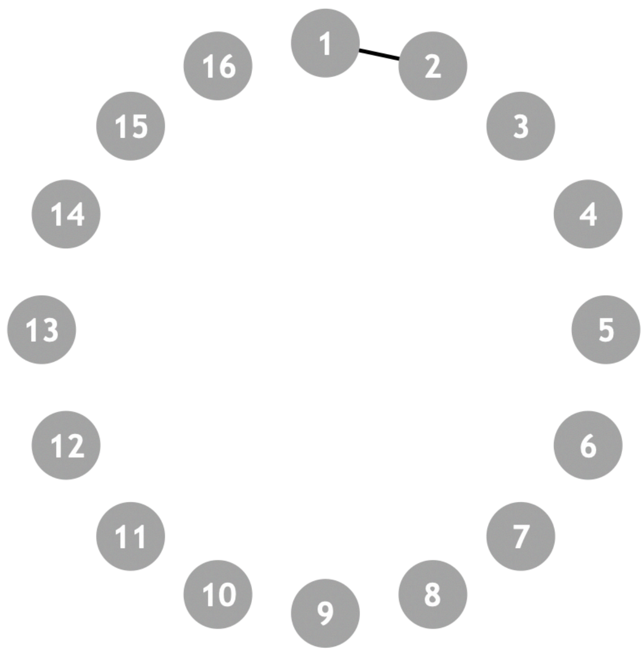


the party algorithm



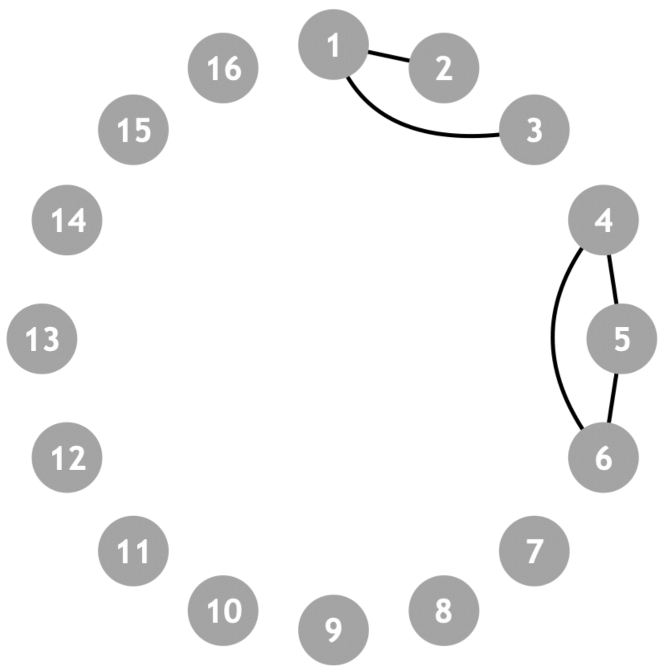
Hagamos un seguimiento de quién habla con quién

the party algorithm



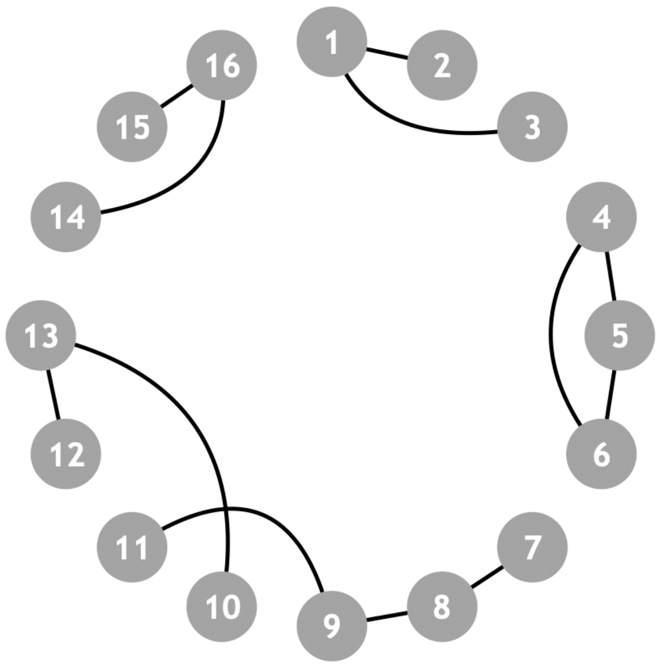
Hagamos un seguimiento de quién habla con quién

the party algorithm



Hagamos un seguimiento de quién habla con quién

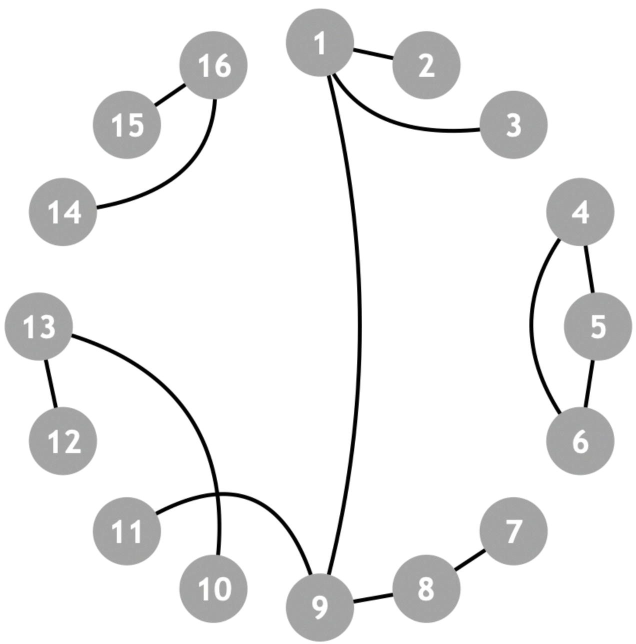
the party algorithm



Hagamos un seguimiento de quién habla con quién

Al principio, la gente habla principalmente con otros que están sentados cerca.

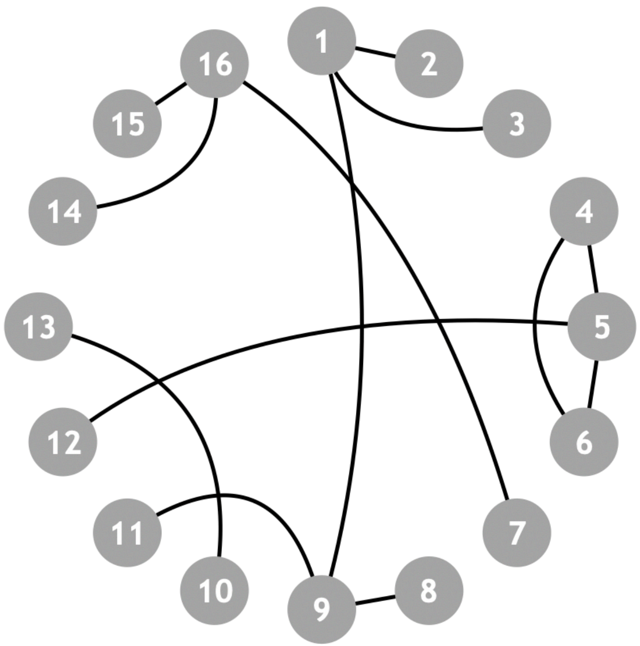
the party algorithm



Hagamos un seguimiento de quién habla con quién

Sin embargo, como la cena avanza hacia las bebidas nocturnas y los invitados se ponen de pie, vemos la reorganización de sus relaciones y se forman nuevos vínculos

the party algorithm

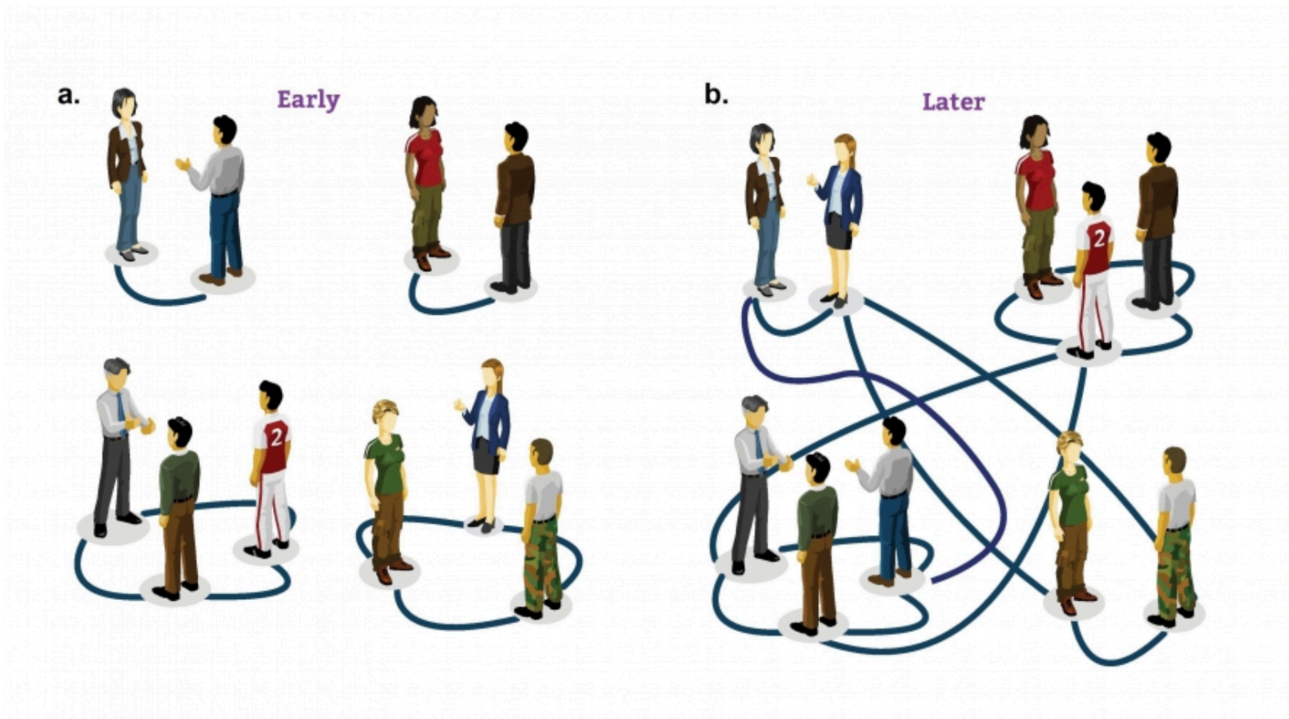


Estas redes no siguen ninguna regla física como en los ejemplos anteriores.

En cambio, son algo aleatorias.
¿No te parece?

Pregunta: ¿Podemos construir un modelo que intente capturar las propiedades de estas redes?

Un algoritmo simple para redes sociales

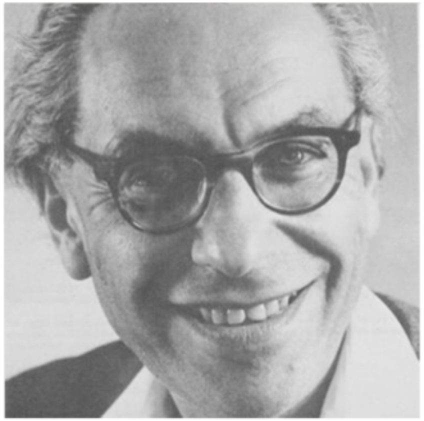


Invitados a una cena se reunen al azar y establecen relaciones sociales.

El modelo de redes aleatorias

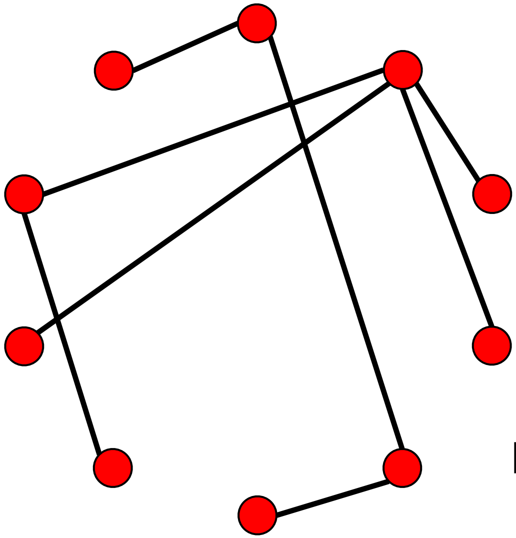
MODELO DE REDES ALEATORIAS

Pául Erdös
(1913-1996)



Alfréd Rényi
(1921-1970)

Erdös-Rényi model (1960)

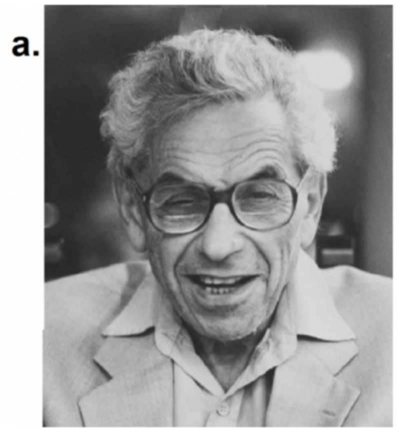


p=1/6 N=10

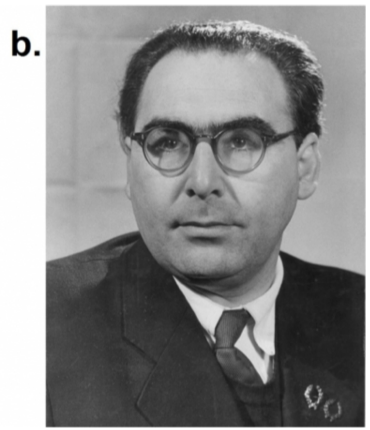
<k> ~ 1.5

Un **gafo aleatorio** es un grafo de **N** nodos donde cada par de nodos está conectado con una probabilidad **p.**

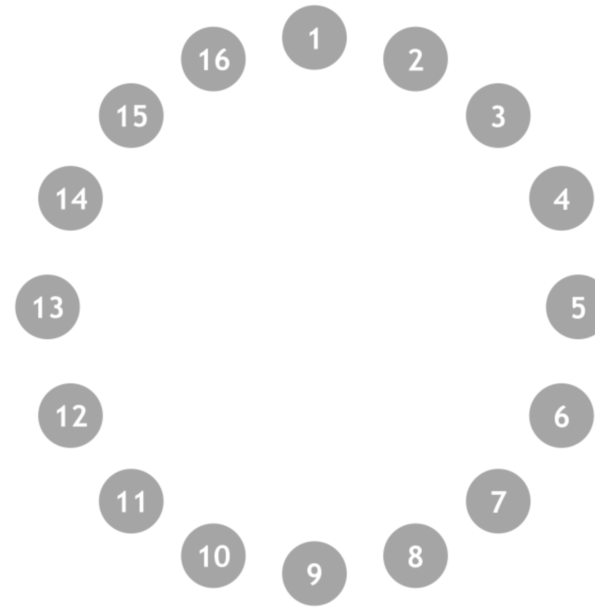
Random Networks



Pál Erdős



Alfréd Rényi

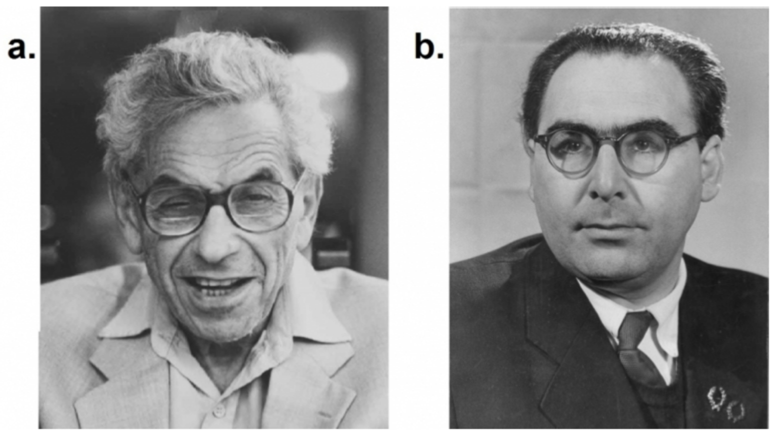


Erdős-Rényi algorithm

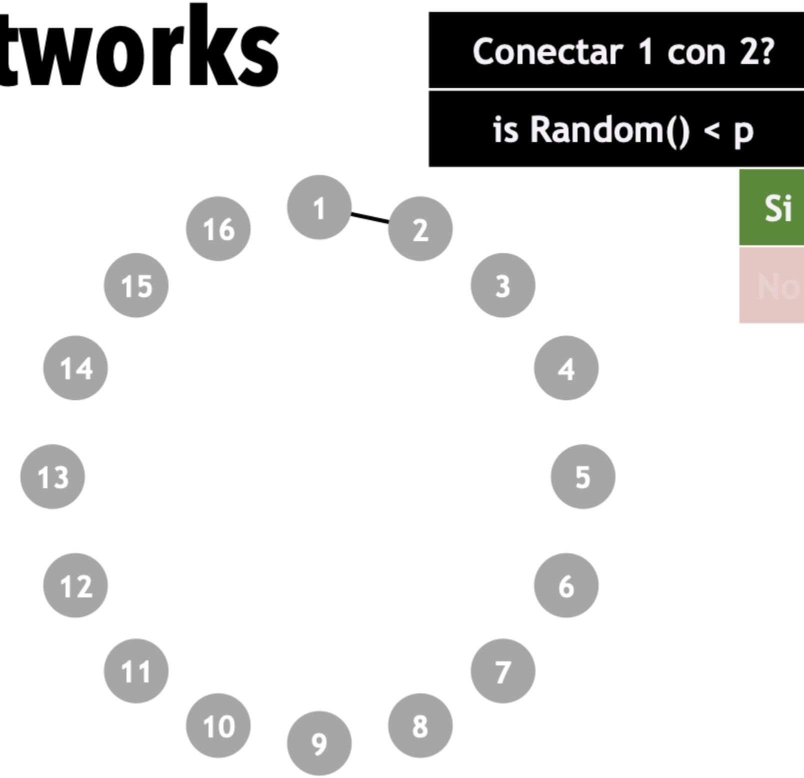
condición inicial Comenzando con N nodos desconectados (L = 0 enlaces)

paso iterativo Para cada par de individuos, establezca un vínculo con probabilidad p

Random Networks



Pál Erdős   Alfréd Rényi



Erdős-Rényi algorithm

condición inicial Comenzando con N nodos desconectados (L = 0 enlaces)

paso iterativo Para cada par de individuos, establezca un vínculo con probabilidad p

Random Networks

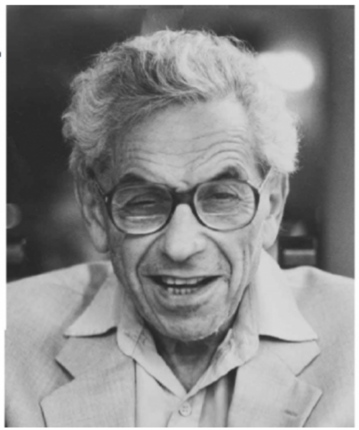
Conectar 1 con 3?

is Random() < p

Si

No

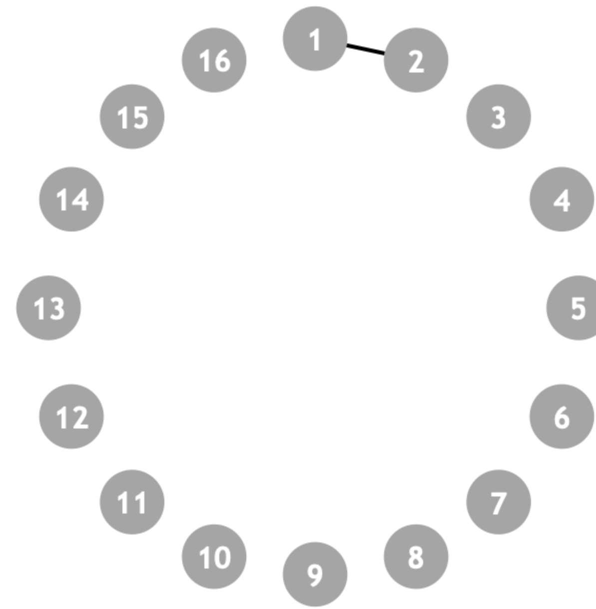
a.



Pál Erdős

b.

Alfréd Rényi



Erdős-Rényi algorithm

condición inicial Comenzando con N nodos desconectados (L = 0 enlaces)

paso iterativo Para cada par de individuos, establezca un vínculo con probabilidad p

Random Networks

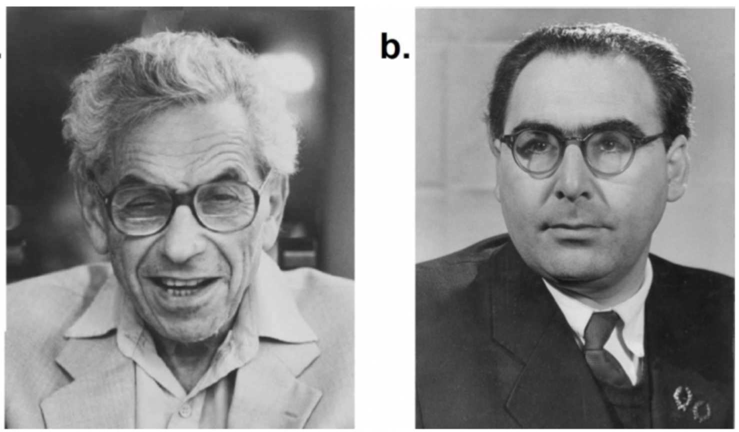
Conectar 1 con 4?

is Random() < p

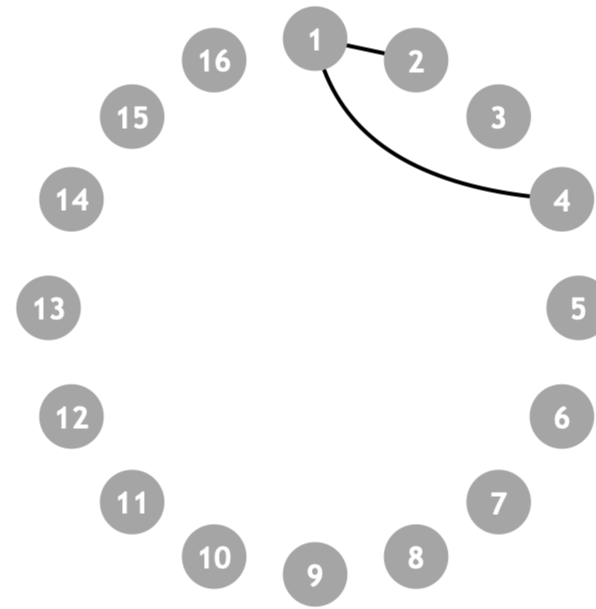
Si

No

a. b.



Pál Erdős   Alfréd Rényi



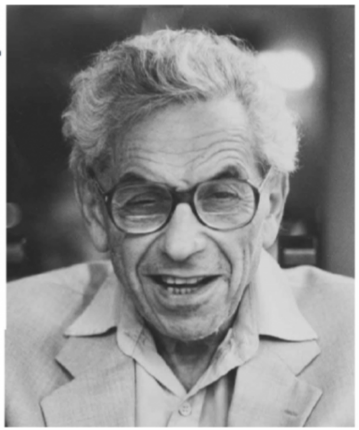
Erdős-Rényi algorithm

condición inicial Comenzando con N nodos desconectados (L = 0 enlaces)

paso iterativo Para cada par de individuos, establezca un vínculo con probabilidad p

Random Networks

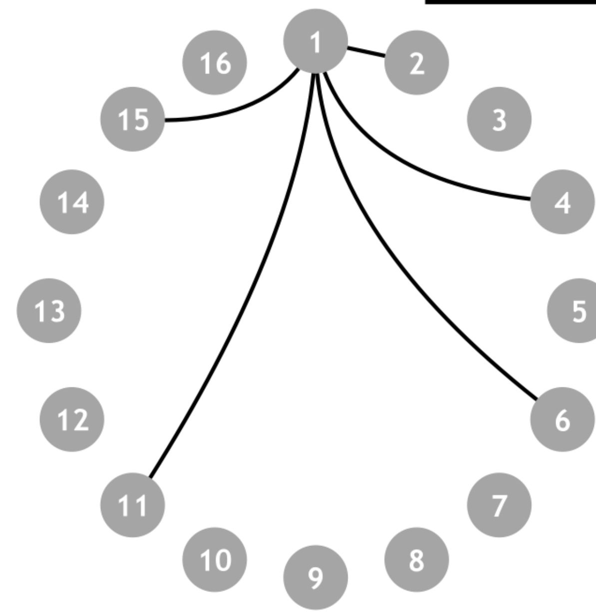
Repite para todas las Posibles conexiones de 1

a. 

Pál Erdős

b. 

Alfréd Rényi



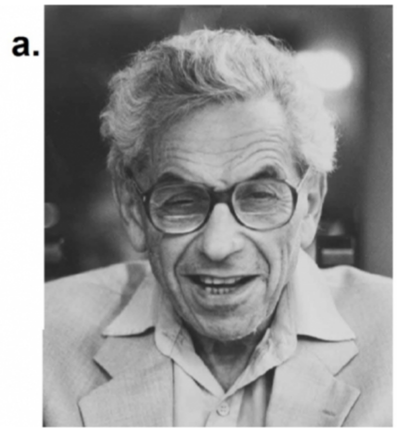
Erdős-Rényi algorithm

condición inicial Comenzando con N nodos desconectados (L = 0 enlaces)

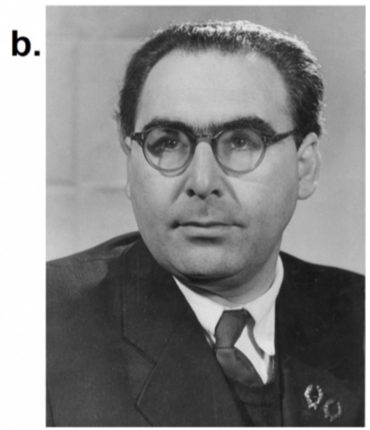
paso iterativo Para cada par de individuos, establezca un vínculo con probabilidad p

Random Networks

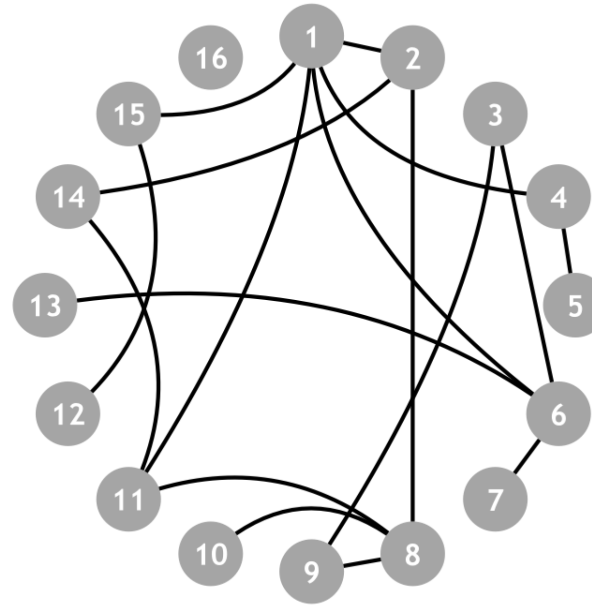
Repite para todos
Los posibles enlaces



Pál Erdős



Alfréd Rényi



Erdős-Rényi algorithm

condición inicial Comenzando con N nodos desconectados (L = 0 enlaces)

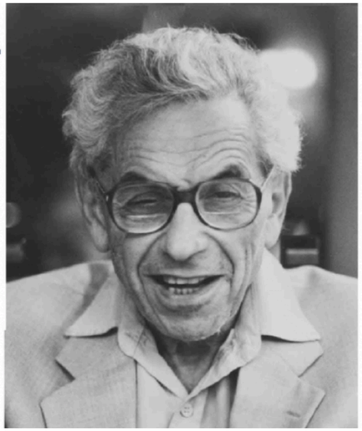
paso iterativo Para cada par de individuos, establezca un vínculo con probabilidad p

El número de enlaces es variable.

MODELO DE REDES ALEATORIAS

Erdős-Rényi algorithm

a.



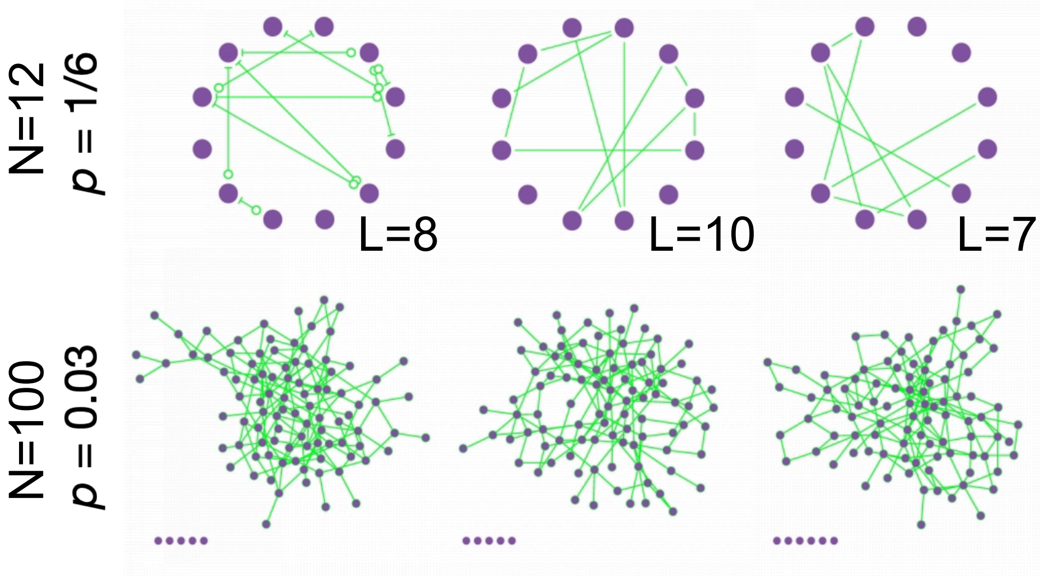
Pál Erdős

b.

Alfréd Rényi

condición inicial Comenzando con N nodos desconectados (L = 0 enlaces)

paso iterativo Para cada par de individuos, establezca un vínculo con probabilidad p



Parece una red real, ¿no es así?

Número de enlaces en una red aleatoria

P(L): la probabilidad de tener exactamente **L** enlaces en una red de **N** nodos y probabilidad **p**:

El número máximo de enlaces en una red de **N** nodos.

$$P(L) = \binom{\binom{N}{2}}{L} p^{L} (1-p)^{\frac{N(N-1)}{2} - L}$$

Número de formas diferentes en que podemos elegir los enlaces L entre todos los enlaces potenciales.

Distribución binomial...

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1},$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

Tutorial Matemático. Distribución Binomial: La línea base

$$p_x = \binom{N}{x} p^x (1-p)^{N-x}$$

Dist. Binomial de "x"

Ejemplo:
N=Enlaces totales potenciales
x=Enlaces de la red
p= probabilidad

$$\langle x \rangle = \sum_{x=0}^{N} x p_x = Np$$

$$< x^2 >= p(1-p)N + p^2 N^2$$

$$\sigma_x = \left(< x^2 > - < x >^2 \right)^{\frac{1}{2}} = [\, p(1-p)N]^{1/2}$$

http://keral2008.blogspot.com/2008/10/derivation-of-mean-and-variance-of.html

MODELO DE REDES ALEATORIAS

P(L): la probabilidad de tener una red de **L** enlaces exactamente

$$P(L) = \binom{\binom{N}{2}}{L} p^{L} (1-p)^{\frac{N(N-1)}{2}-L}$$

•El número promedio de links <L> en un grafo aleatorio

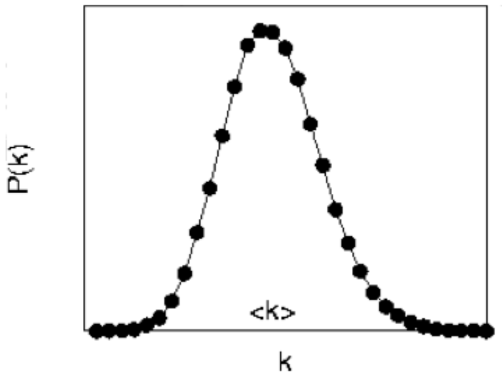
$$< L > = \sum_{L=0}^{\frac{N(N-1)}{2}} L P(L) = p\frac{N(N-1)}{2} \longrightarrow < k > = 2L/N = p(N-1)$$

•La varianza

$$\sigma^{2} = p(1-p)\frac{N(N-1)}{2}$$

Distribución de grado de una red aleatoria

DISTRIBUCION DE GRADO DE UNA RED ALEATORIA



$$P(k) = \binom{N-1}{k} p^k (1-p)^{(N-1)-k}$$

Selecciona k nodos de un conjunto de N-1 nodos

Probabilidad de tener *k* enlaces

Probabilidiad de perder N-1-k enlaces

$$<k> = p(N-1) \qquad \sigma_k^2 = p(1-p)(N-1)$$

$$\frac{\sigma_k}{<k>} = \left[\frac{1-p}{p}\frac{1}{(N-1)}\right]^{1/2} \approx \frac{1}{(N-1)^{1/2}}$$

A medida que la red aumenta de tamaño, la distribución se vuelve cada vez más estrecha: estamos cada vez más seguros de que el grado de un nodo se encuentra cerca de <k> (valores homogeneos).

DISTRIBUCION DE GRADO DE UNA RED ALEATORIA

$$P(k) = \binom{N-1}{k} p^k (1-p)^{(N-1)-k} \qquad <k> = p(N-1) \qquad p = \frac{<k>}{(N-1)}$$

Para *N* grandes y *k* pequeños, Podemos usar las siguientes aproximaciones (N>>k):

$$\binom{N-1}{k} = \frac{(N-1)!}{k!(N-1-k)!} = \frac{(N-1)(N-1-1)(N-1-2)\ldots(N-1-k+1)(N-1-k)!}{k!(N-1-k)!} = \frac{(N-1)^k}{k!}$$

$$\ln[(1-p)^{(N-1)-k}] = (N-1-k)\ln(1-\frac{<k>}{N-1}) = -(N-1-k)\frac{<k>}{N-1} = -<k>(1-\frac{k}{N-1}) \cong -<k>$$

$$(1-p)^{(N-1)-k} = e^{-<k>} \qquad \ln(1+x) = \sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}x^n = x - \frac{x^2}{2} + \frac{x^3}{3} - \ldots \quad \text{for} \quad |x| \leq 1$$

$$\boxed{P(k) = \binom{N-1}{k} p^k (1-p)^{(N-1)-k} = \frac{(N-1)^k}{k!} p^k e^{-<k>} = \frac{(N-1)^k}{k!}\left(\frac{<k>}{N-1}\right)^k e^{-<k>} = e^{-<k>}\frac{<k>^k}{k!}}$$

DISTRIBUCIÓN DE GRADO DE POISSON

$$P(k) = \binom{N-1}{k} p^k (1-p)^{(N-1)-k}$$

$$<k> = p(N-1)$$

$$p = \frac{<k>}{(N-1)}$$

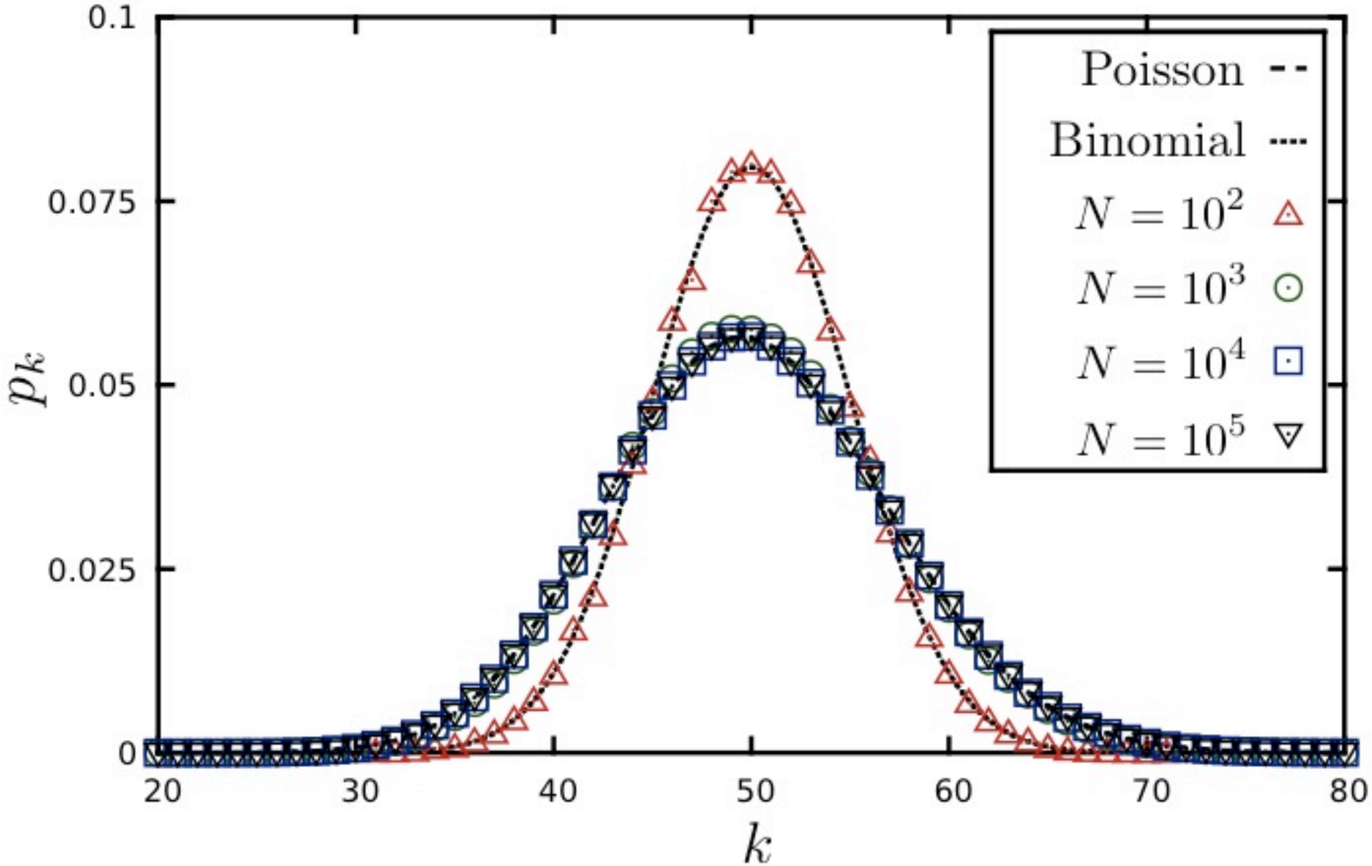
Para *N* grandes y *k* pequeños, llegamos a la distribución de Poisson:

$$P(k) = e^{-<k>} \frac{<k>^k}{k!}$$

DISTRIBUCION DE GRADO DE UNA RED ALEATORIA

$<k>=50$

$$P(k) = e^{-<k>} \frac{<k>^k}{k!}$$



K

P(k): k<<N
Indistinguibles

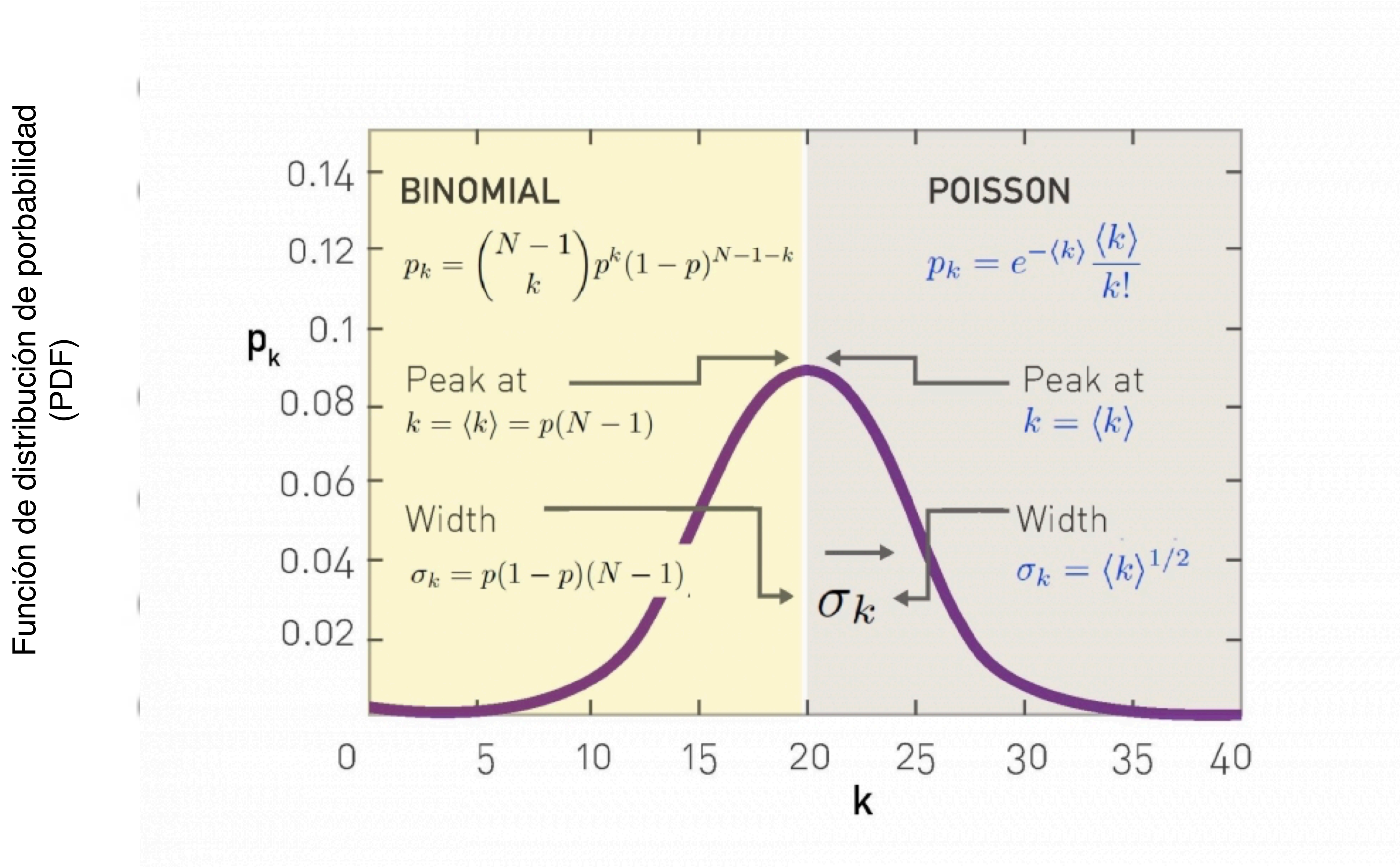
DISTRIBUCION DE GRADO DE UNA RED ALEATORIA

Resultado exacto

- Distribución Binomial -

Límite para N grande

- Distribución de Poisson -



Las redes reales **no** son Poisson

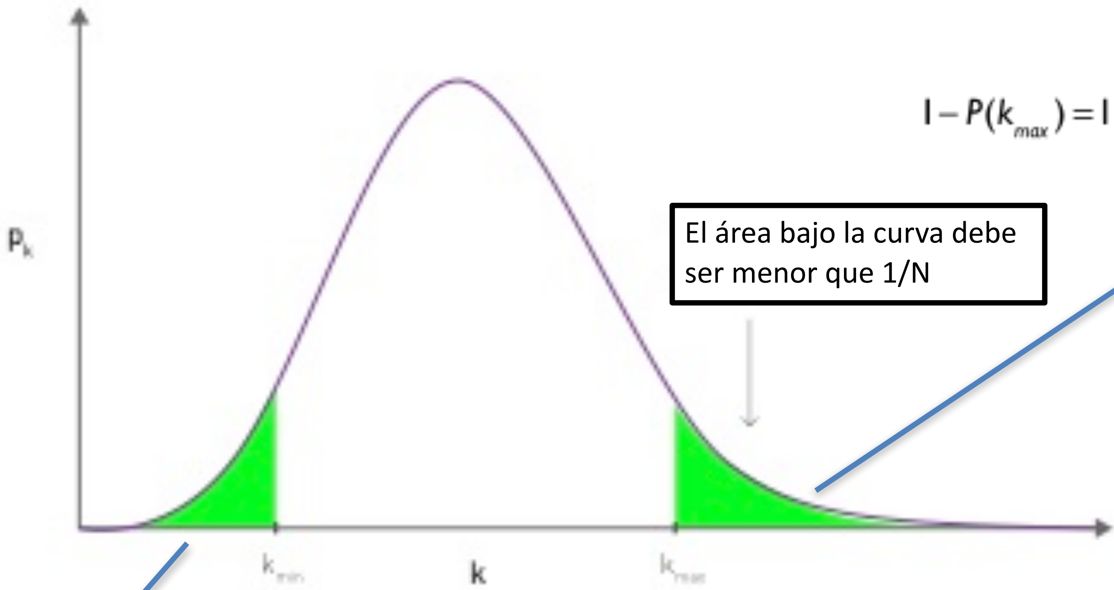
Grado máximo y mínimo

$$P(k) = e^{-<k>} \frac{<k>^k}{k!}$$

Ej: $<k>=1{,}000$, $N=10^9$

$$N\left[1 - P(k_{max})\right] \approx 1.$$

$$1 - P(k_{max}) = 1 - e^{-\langle k \rangle} \sum_{k=0}^{k_{max}} \frac{\langle k \rangle^k}{k!} = e^{-\langle k \rangle} \sum_{k=k_{max}+1}^{\infty} \frac{\langle k \rangle^k}{k!} \approx e^{-\langle k \rangle} \frac{\langle k \rangle^{k_{max}+1}}{(k_{max}+1)!},$$



El área bajo la curva debe ser menor que 1/N

$k_{max} = 1{,}185$

$$NP(k_{min}) \approx 1.$$

$$P(k_{min}) = e^{-\langle k \rangle} \sum_{k=0}^{k_{min}} \frac{\langle k \rangle^k}{k!}.$$

$k_{min}=816$

$$<k> \pm \sigma_k \qquad \sigma_k = <k>^{1/2}$$

$$\sigma_k = 31.62.$$

NO HAY “OUTLIERS” EN UNA SOCIEDAD ALEATORIA

Supongamos que una persona típica conoce aproximadamente 1,000 individuos por su nombre de pila

$$P(k) = e^{-<k>} \frac{<k>^k}{k!}$$

→El individuo más conectado tiene grado k_{max}~1,185

$$1 - P(k_{max}) = 1 - e^{-\langle k \rangle} \sum_{k=0}^{k_{max}} \frac{\langle k \rangle^k}{k!} = e^{-\langle k \rangle} \sum_{k=k_{max}+1}^{\infty} \frac{\langle k \rangle^k}{k!} \approx e^{-\langle k \rangle} \frac{\langle k \rangle^{k_{max}+1}}{(k_{max}+1)!},$$

→El individuo menos conectado tiene grado k_{min} ~ 816

$$P(k_{min}) = e^{-\langle k \rangle} \sum_{k=0}^{k_{min}} \frac{\langle k \rangle^k}{k!}.$$

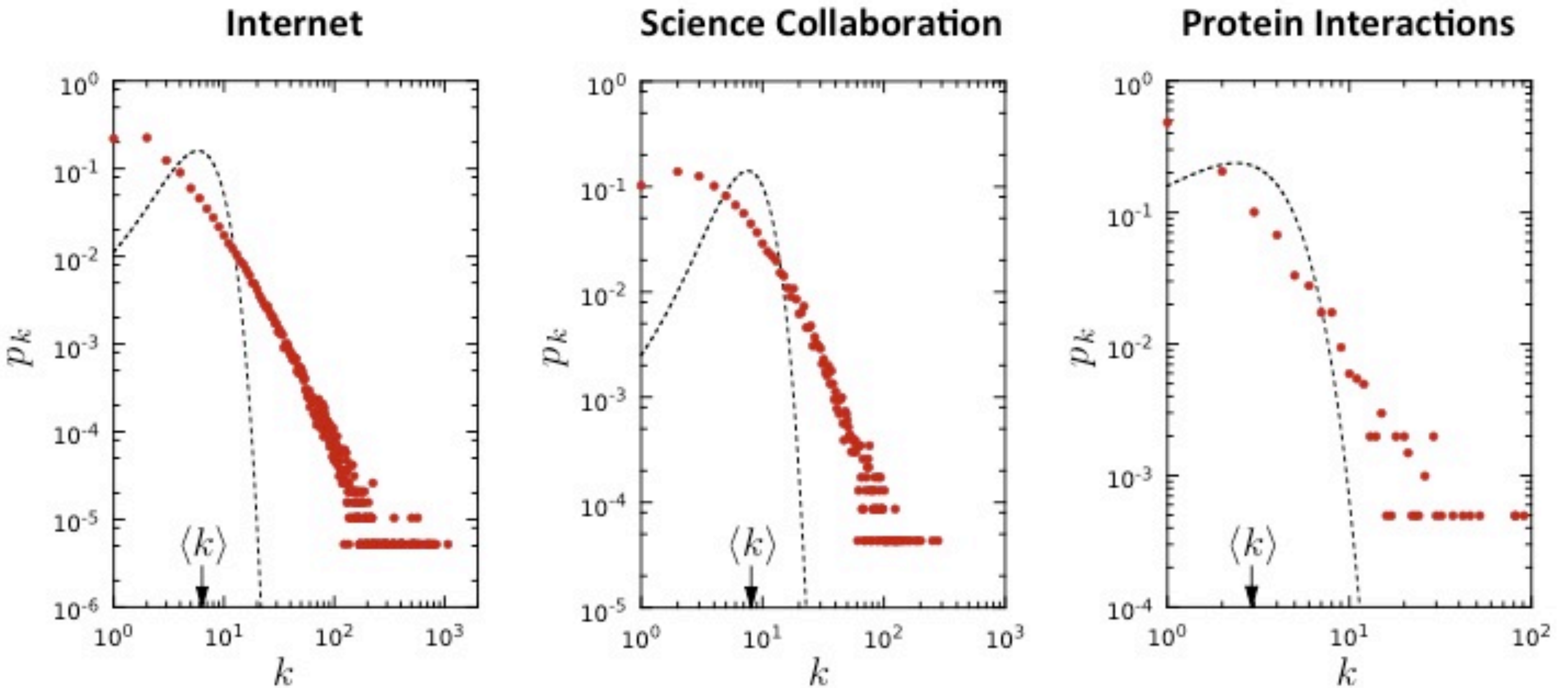
La probabilidad de encontrar un individuo con grado k>2,000 is 10^{-27}. Por lo tanto, la posibilidad de encontrar a un individuo con 2,000 conocidos es tan pequeña que tales nodos son virtualmente inexistentes en una sociedad aleatoria.

→una sociedad aleatoria consistiría principalmente en individuos promedio, con todos con aproximadamente el mismo número de amigos.

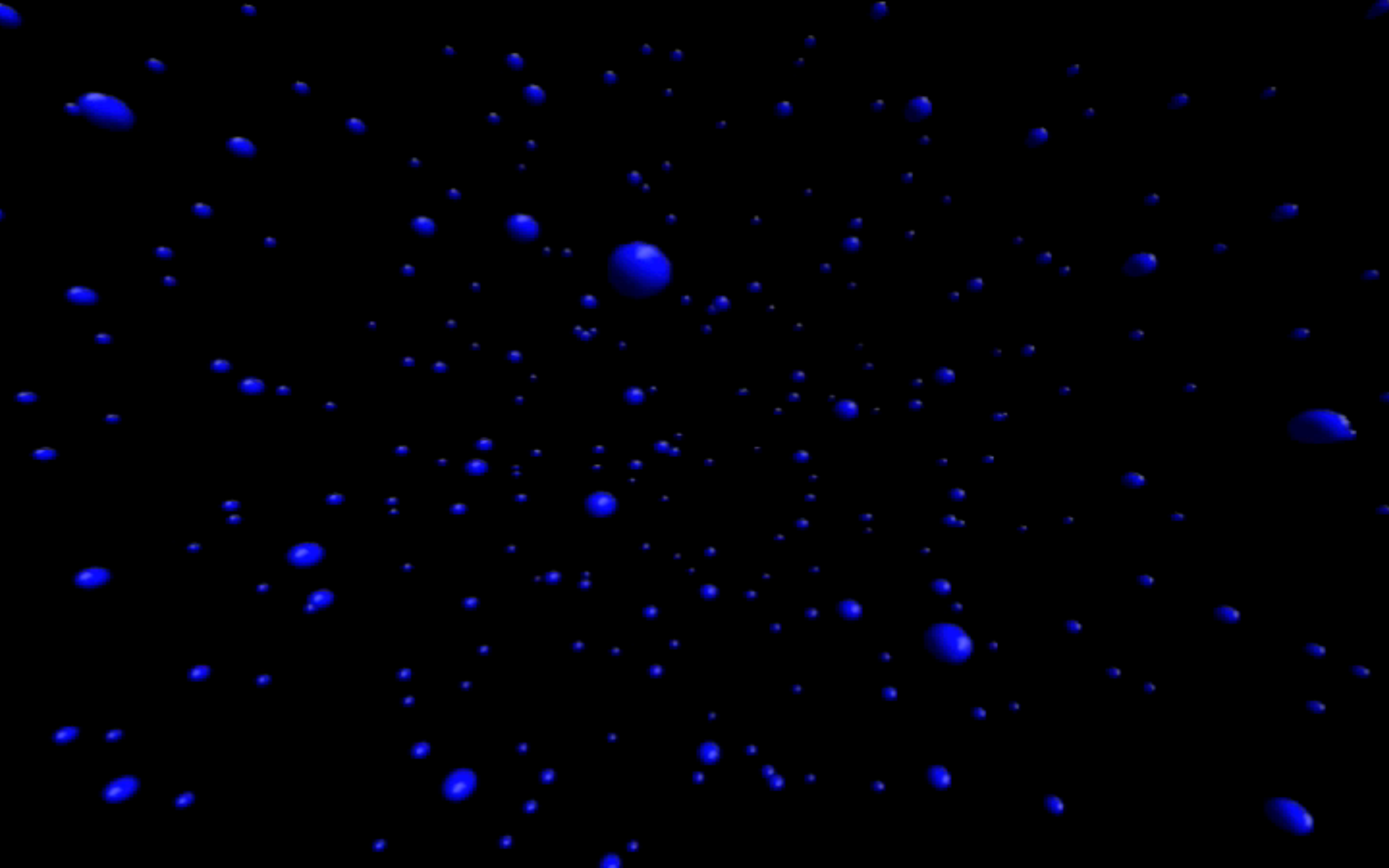
→Carecería de valores atípicos, individuos que son muy populares o solitarios.

ENFRENTANDO LA REALIDAD: Distribución en grado de redes reales.

$$P(k) = e^{-<k>} \frac{<k>^k}{k!}$$

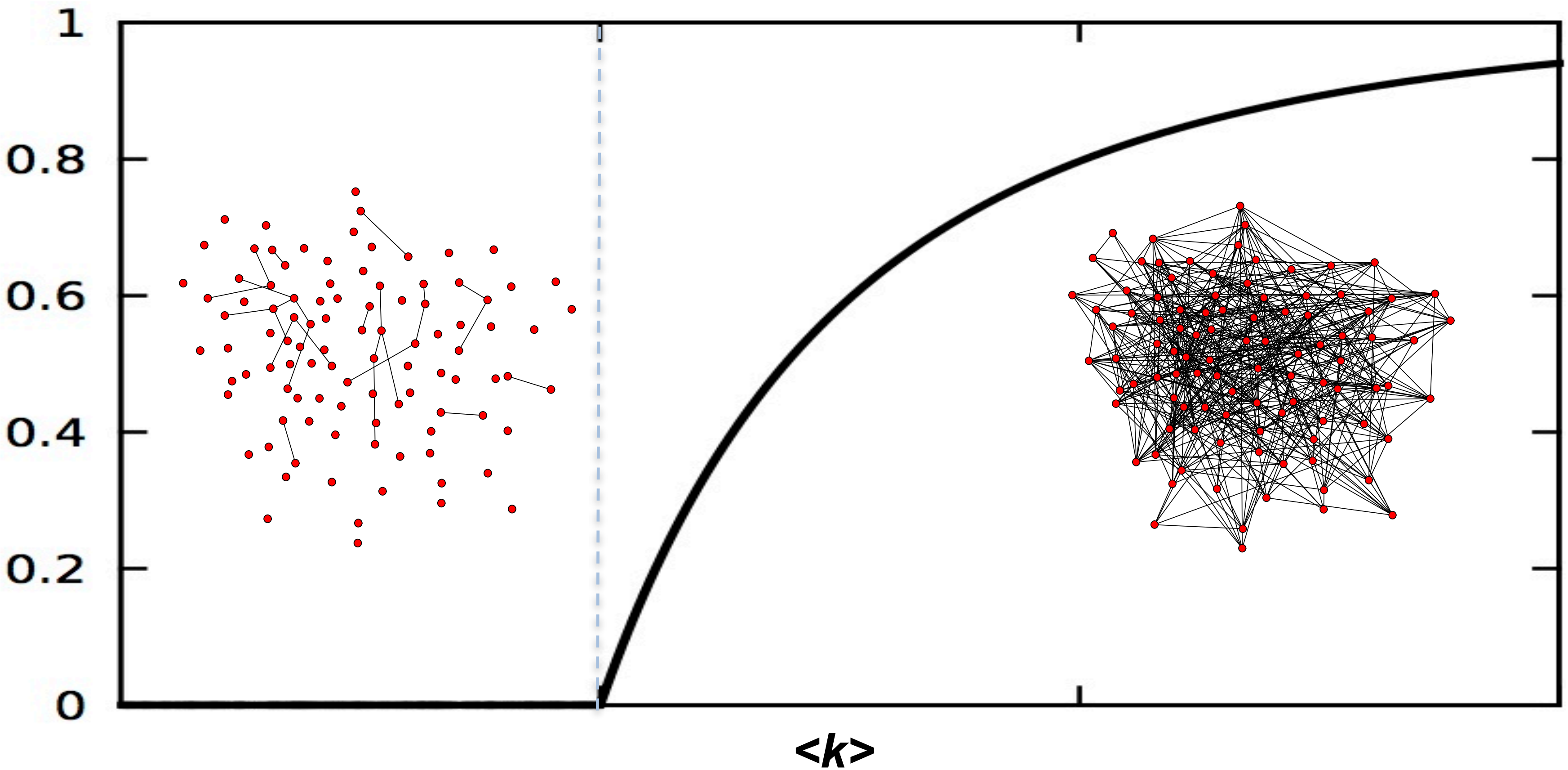


La evolución de una red aleatoria.



Evolución de una red aleatoria

Nodos desconectados ➔ **Red**.

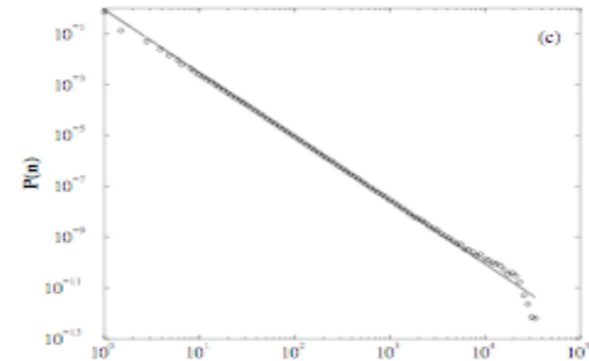


¿Cómo se produce esta transición?

DISTRIBUCION DEL TAMAÑO DEL GRUPO (CLUSTER)

Probabilidad de que un nodo seleccionado al azar pertenezca a un grupo de tamaño *s*:

$$p(s) = \frac{e^{-\langle k\rangle s}(\langle k\rangle s)^{s-1}}{s!} \qquad \langle k\rangle^{s-1} = \exp[(s-1)\ln\langle k\rangle]$$

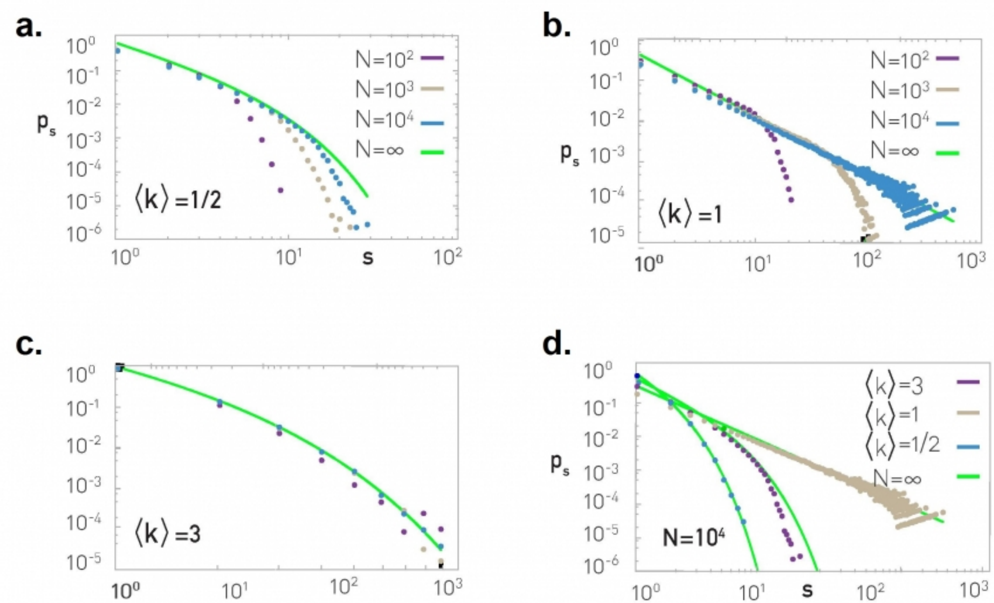


$$p(s) = \frac{s^{s-1}}{s!} e^{-\langle k\rangle s+(s-1)\ln\langle k\rangle} \qquad s! = \sqrt{2\pi s}\left(\frac{s}{e}\right)^{s}$$

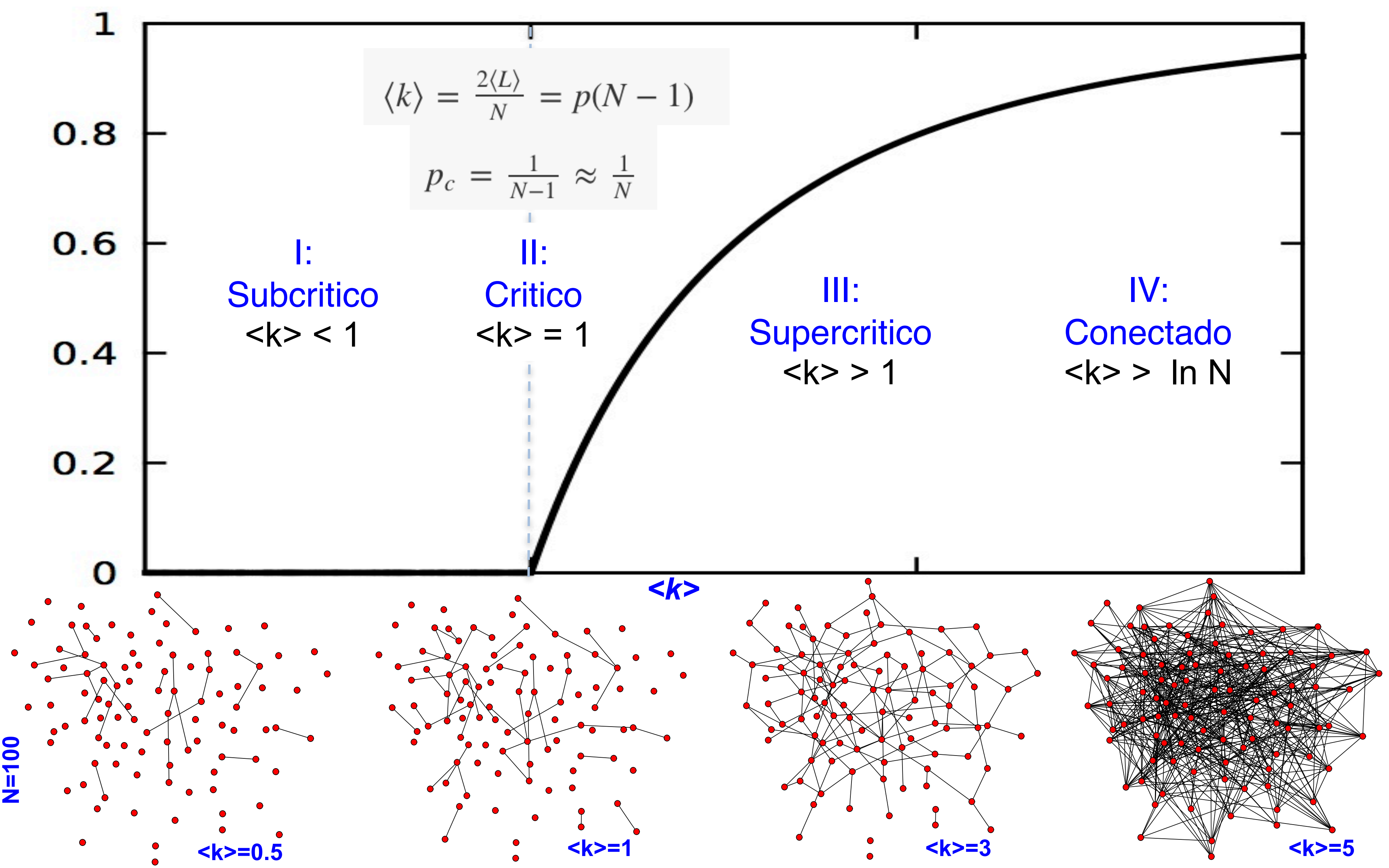
$$p(s) \sim s^{-3/2} e^{-(\langle k\rangle -1)s+(s-1)\ln\langle k\rangle}$$

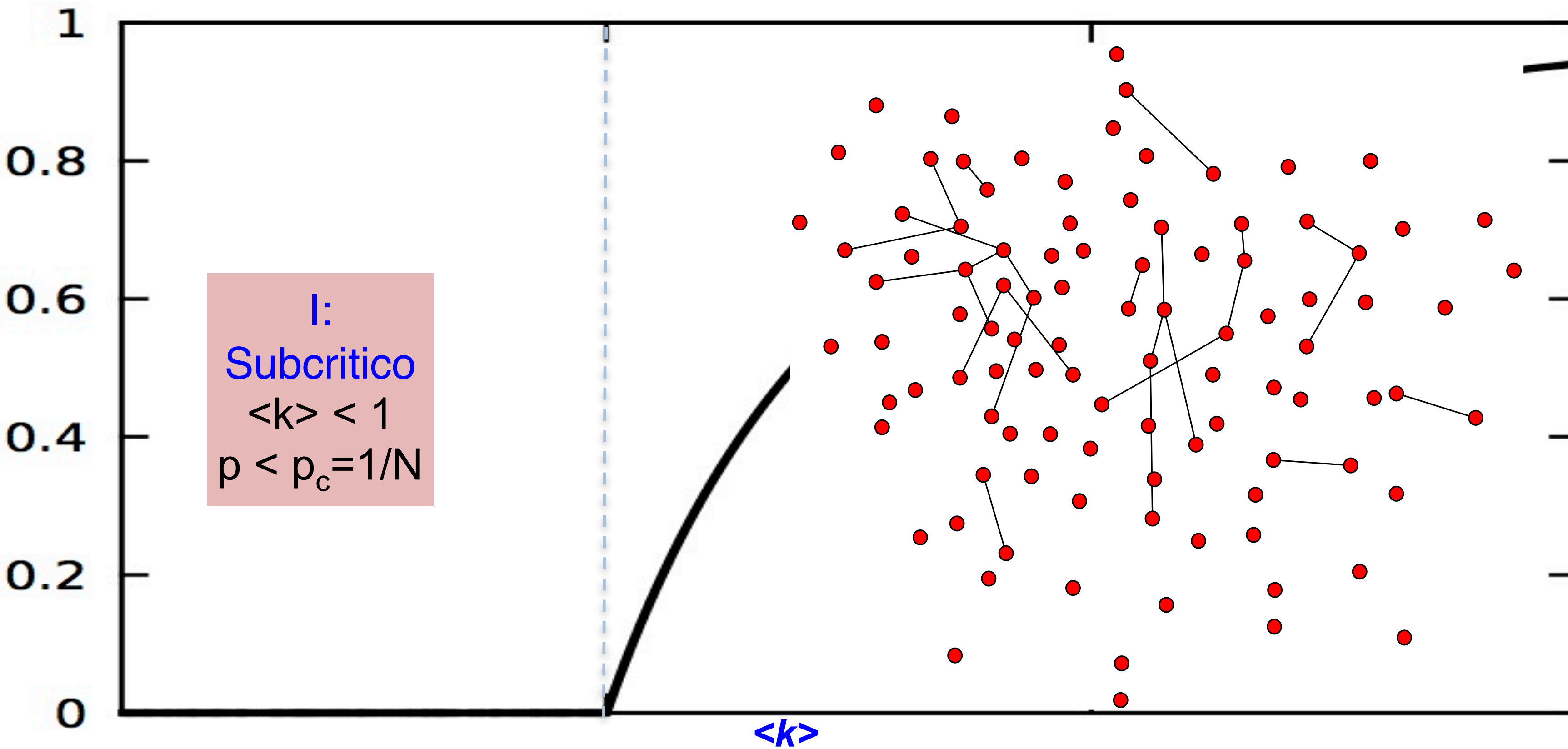
At the critical point <k>=1

$$p(s) \sim s^{-3/2}$$



Derivation in Newman, 2010

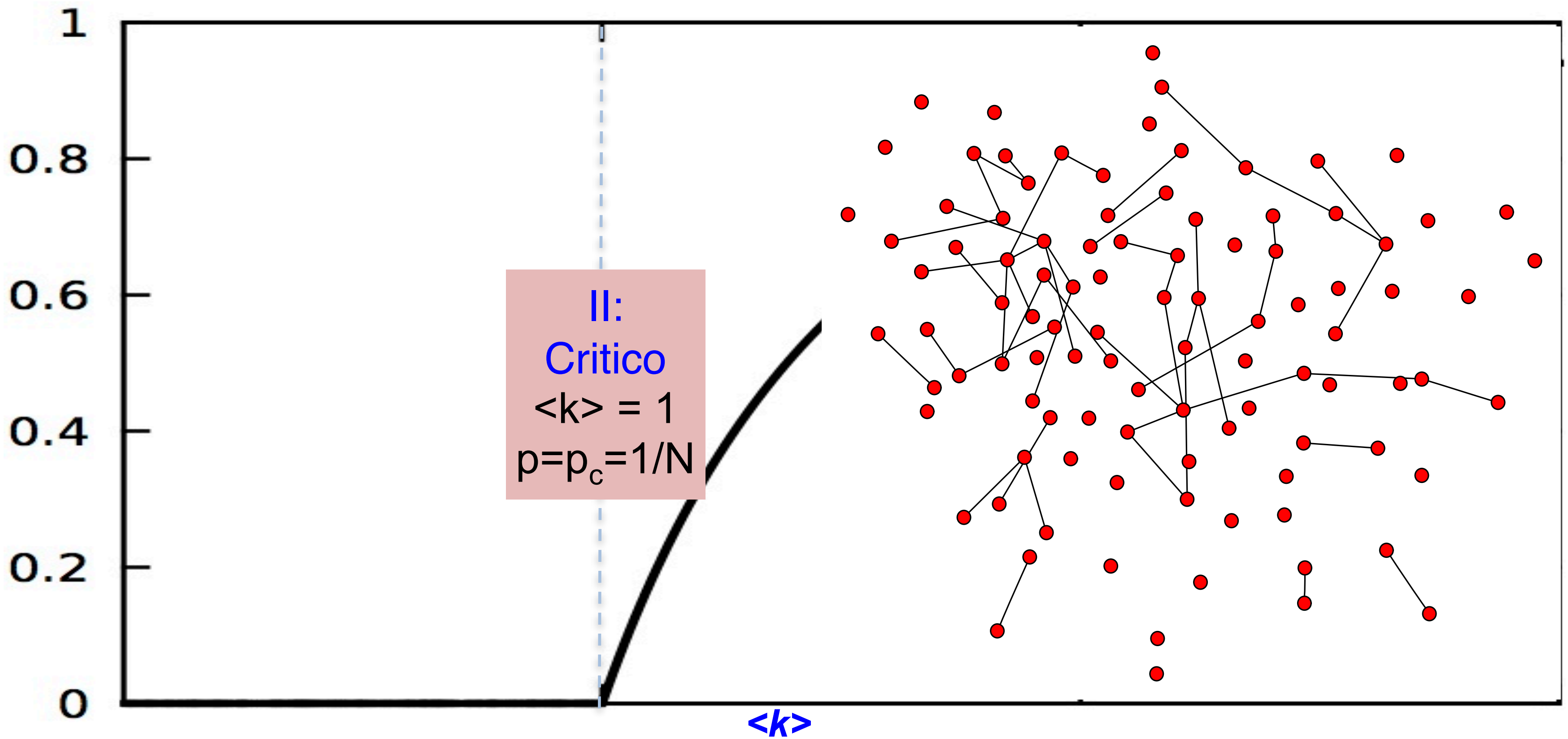

$\langle k \rangle = \frac{2\langle L \rangle}{N} = p(N-1)$
$p_c = \frac{1}{N-1} \approx \frac{1}{N}$
I:
Subcritico
<k> < 1
II:
Critico
<k> = 1
III:
Supercritico
<k> > 1
IV:
Conectado
<k> > ln N
1
0.8
0.6
0.4
0.2
0
<k>
N=100
<k>=0.5
<k>=1
<k>=3
<k>=5



No hay componente gigante

$$p(s) \sim s^{-3/2} e^{-(\langle k \rangle - 1)s + (s-1)\ln\langle k \rangle}$$

N-L clusters aislados, la distribución del tamaño del cluster es exponencial
El cluster mas grande es un árbol, su tamaño es ~ *ln N*



Una sola componente gigante: $N_G \sim N^{2/3}$

→Contiene una decreciente fracción de todos los nodos, $N_G/N \sim N^{-1/3}$

→Componentes pequeños son árboles, GC tiene loops

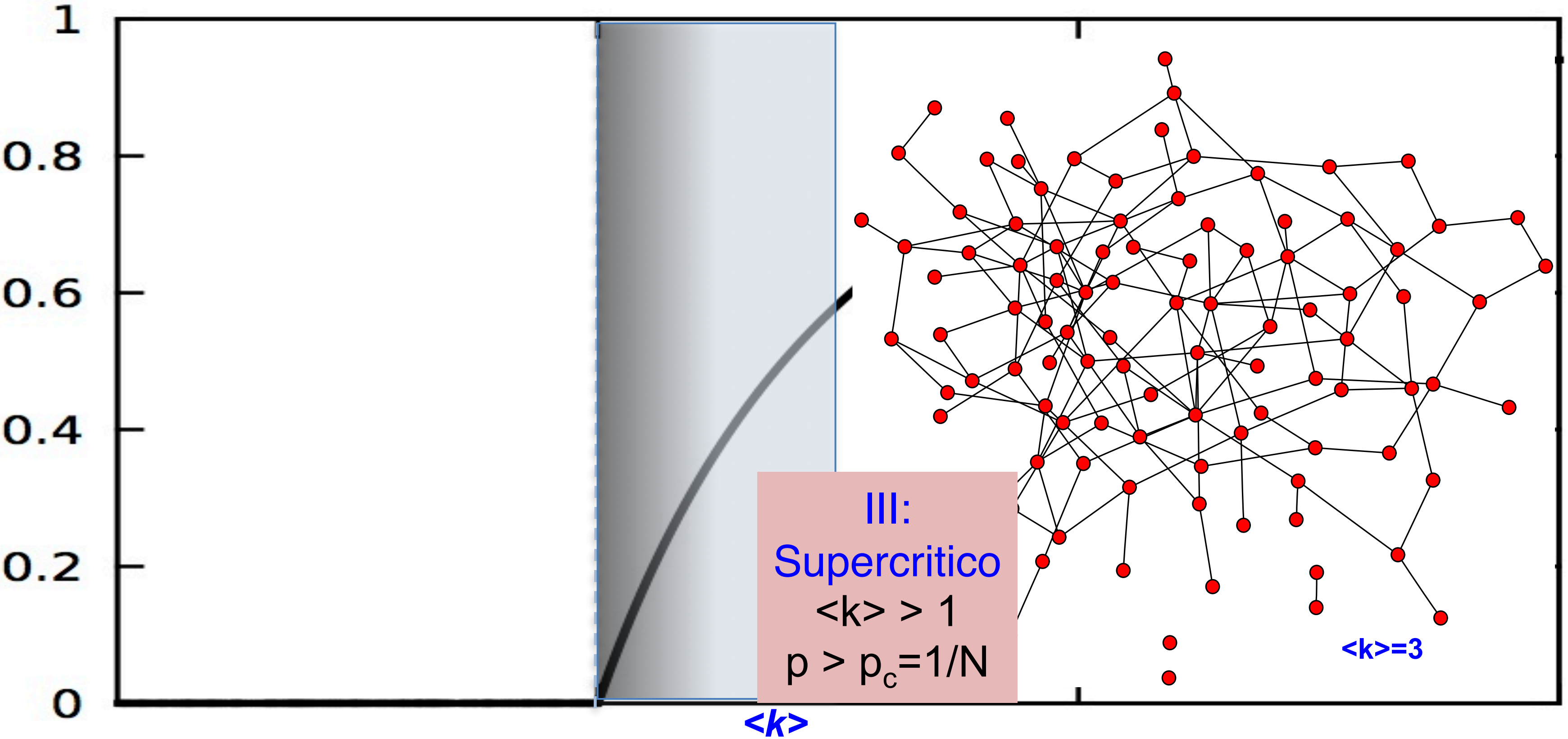
Distribución del tamaño de cluster: $p(s) \sim s^{-3/2}$

(Power-law/ley de potencia)

Un salto en el tamaño del cluster:

N=1,000 → ln N~ 6.9; $N^{2/3} \sim 95$

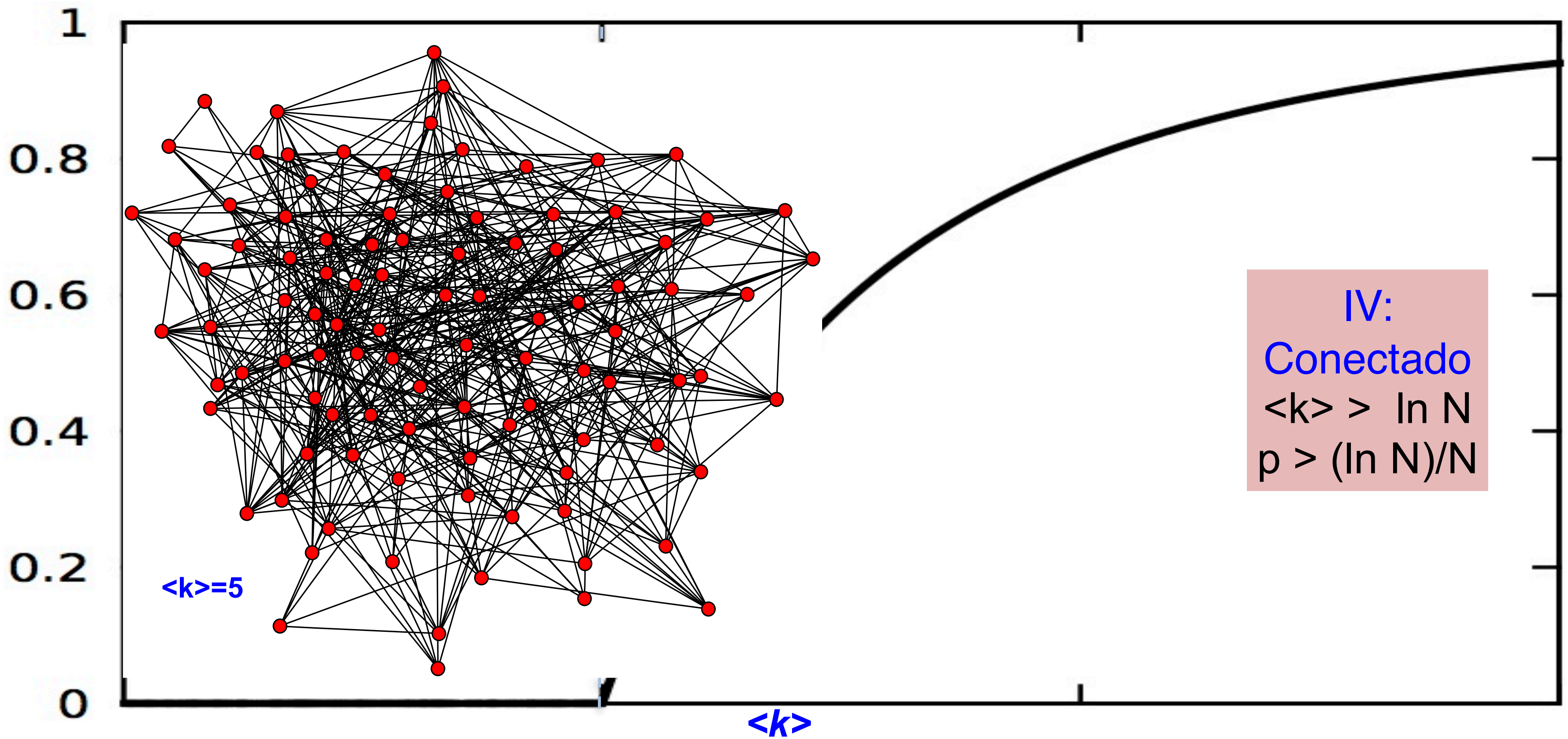
N=7 10^9 → ln N~ 22; $N^{2/3} \sim 3{,}659{,}250$



Una sola componente gigante: $N_G \sim (p-p_c)N$

→GC tiene loops.

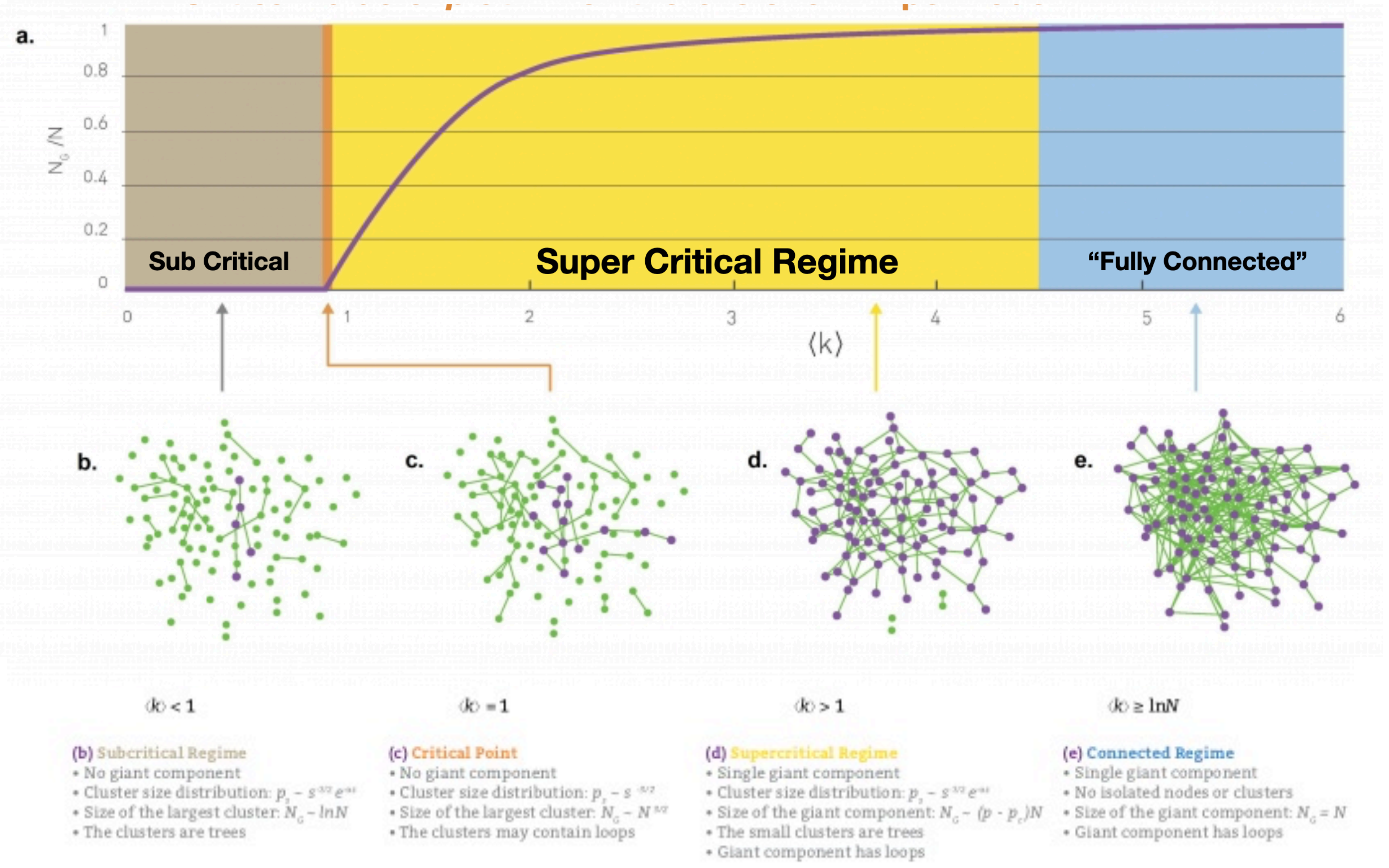
Distribución de tamaños de clusters: exponencial $p(s) \sim s^{-3/2} e^{-(\langle k \rangle -1)s+(s-1)\ln\langle k \rangle}$



Solo un cluster: N_G=N

→GC es densa.

Distribución de tamaño de clusters: Ninguna



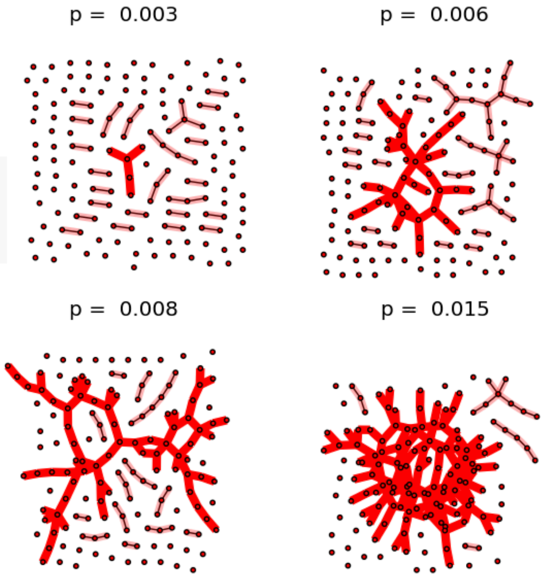
a.
1
0.8
0.6
0.4
0.2
0
N_G/N
Sub Critical
Super Critical Regime
“Fully Connected”
0
1
2
3
4
5
6
⟨k⟩
b.
c.
d.
e.
$\langle k \rangle < 1$
$\langle k \rangle = 1$
$\langle k \rangle > 1$
$\langle k \rangle \geq \ln N$
(b) Subcritical Regime
• No giant component
• Cluster size distribution: $p_s \sim s^{-3/2} e^{-\alpha s}$
• Size of the largest cluster: $N_G \sim \ln N$
• The clusters are trees
(c) Critical Point
• No giant component
• Cluster size distribution: $p_s \sim s^{-3/2}$
• Size of the largest cluster: $N_G \sim N^{2/3}$
• The clusters may contain loops
(d) Supercritical Regime
• Single giant component
• Cluster size distribution: $p_s \sim s^{-3/2} e^{-\alpha s}$
• Size of the giant component: $N_G \sim (p - p_c)N$
• The small clusters are trees
• Giant component has loops
(e) Connected Regime
• Single giant component
• No isolated nodes or clusters
• Size of the giant component: $N_G = N$
• Giant component has loops

Evolución de una red aleatoria

Nodos desconectados ➔ **Red.**

$$\langle k \rangle = \frac{2\langle L \rangle}{N} = p(N-1)$$

$$p_c = \frac{1}{N-1} \approx \frac{1}{N}$$



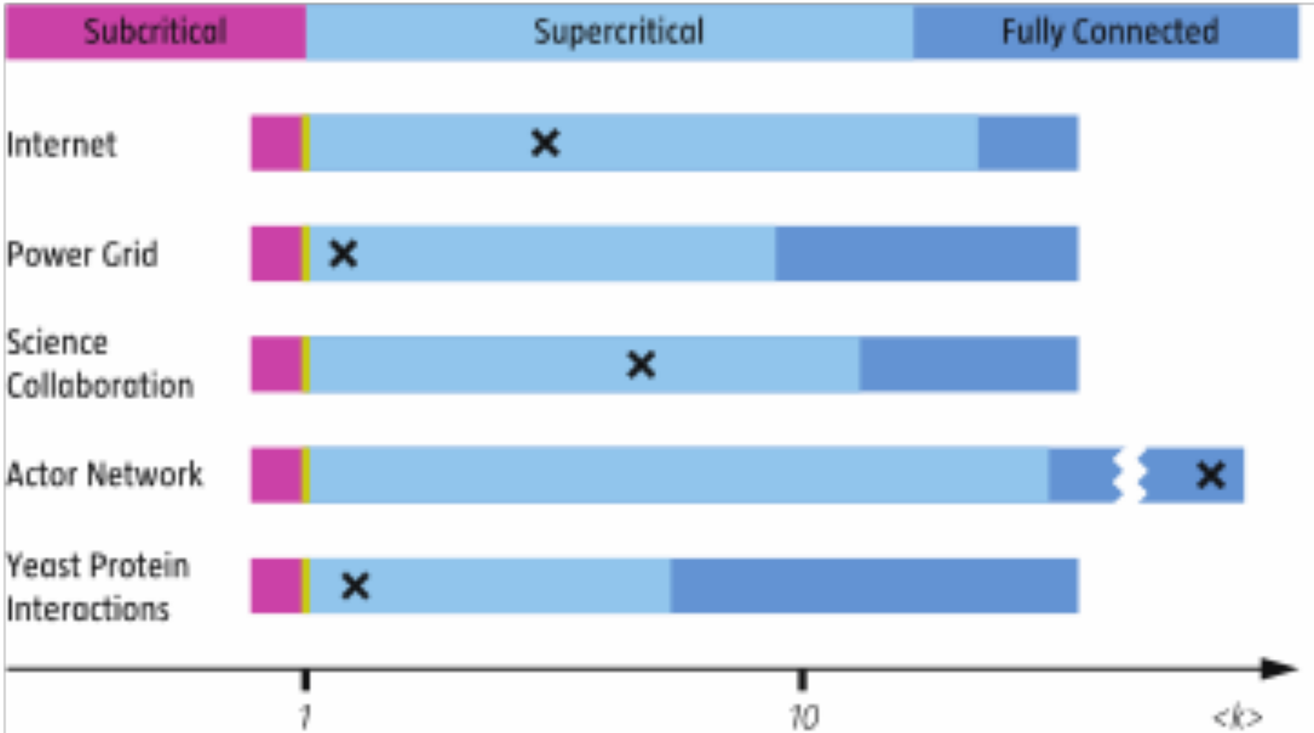
$<k_c>$=1/N * (N-1)=1, para N grande
(Erdos y Renyi, 1959)

El hecho de que al menos un enlace por nodo sea ***necesario*** para tener un componente gigante no es inesperado. De hecho, para que exista un componente gigante, cada uno de sus nodos debe estar vinculado al menos a otro nodo.

Es algo inesperado, sin embargo, que un enlace es **suficiente** para el surgimiento de un componente gigante.

Es igualmente interesante que la aparición del componente gigante no sea gradual, pero sigue lo que los físicos llaman una **transición de fase** de segundo orden en <k> = 1.

Las redes reales son supercríticas.



Network	N	L	‹K›	lnN
Internet	192,244	609,066	6.34	12.17
Power Grid	4,941	6,594	2.67	8.51
Science Collaboration	23,133	94,437	8.08	10.05
Actor Network	702,388	29,397,908	83.71	13.46
Protein Interactions	2,018	2,930	2.90	7.61

Red conectada

$\langle k \rangle > \ln N$

Preguntas básicas

- Qué representa la distribución de grado en una red y por qué es un concepto importante en análisis de redes?
- De qué manera se relaciona el concepto de centralidad de grado con distribución de grado de una red?
- Explica el concepto de una red aleatoria y cómo difiere de una red real.
- Considera una red aleatoria de $N = 2{,}000$. Cuál es el numero estimado de links a una probabilidad $p = 10^{-3}$? Y qué pasa con esa red si agregas el número de nodos a $N = 3{,}000$? Encuentra la probabilidad p_c de manera que la red se encuentre en su punto critico.